

Skimming for Context

Master's Thesis in Computer Science by
Ole Torp Lassen (DIKU)

Dept. of Computer Science, University of Copenhagen
September 2005

advisors:

Neil D. Jones (DIKU)
Peter R. Skadhauge (CBS)

ensor:

Henning Christiansen (RUC)

1 INTRODUCTION	6
1.1 Basic Idea.....	6
1.2 Processing natural language	7
1.3 Words in context.....	10
1.4 Aim of the project.	11
2 NATURAL LANGUAGE PROCESSING (NLP).....	13
2.2 Context.....	14
2.3 Ambiguity	15
2.4 Applications of skimming.....	16
3 THEORETICAL ANALYSIS	17
3.1 Conventions and decisions	17
3.1.1 Typography.....	17
3.1.2 Technical terms.....	17
3.1.3 A word on illustrations, figures and examples	19
3.2 Definitions and restrictions.....	19
3.2.1 Definition of context.....	19
3.2.2 Restrictions on data.....	19
3.2.3 Restrictions on analytical levels	20
3.2.4 Restrictions on concepts	20
3.2.5 Restrictions on accuracy	20
3.2.6 English as data language.....	21
3.2.7 The experimental tasks	21
3.3 Semantic theories and philosophies	21
3.3.1 Lexical semantics.....	23
3.3.2 Synonymy and polysemy.....	24
3.3.3 Lexical relations.....	28
3.3.4 Pragmatics - the Cooperative Principle	32
4 THE MACHINE-READABLE DICTIONARY	36
4.1 Structure of the database	37

4.1.1 An illustrated example.....	38
4.2 Some remarks to the use of WordNet.....	41
5 THE EXPERIMENTAL CORPUS	43
5.1 Civilization III – advantages and problems	44
5.2 The tagger – TOSCA-ICLE.....	45
5.2.1 Pre-processing.....	46
5.3 Solving the problems	46
6 PUTTING IT ALL TOGETHER.	51
6.1 The word game – an illustrated example.....	52
6.2 Problem context	55
6.2.1 Sets of the dictionary :	55
6.3 The task.....	57
6.3.1 Implications	57
6.3.2 On Goodness.....	59
6.4 My solution	59
6.4.1 The CHUNK.....	59
6.4.3 Interpreting the CHUNK	61
6.5 Partial interpretations:.....	63
6.5.1 The algorithm, informally and explicitly.....	64
6.5.2 From non-determinism to determinism	65
6.5.3 How partial interpretations relate to complete interpretations	66
6.5.4 How to compare interpretations.....	68
6.5.5 Two scoring schemes.....	69
6.6 Remarks.....	72
6.6.1 Relations between lexemes.....	72
6.6.2 Complexity analysis.....	74
7 OUTPUT – EXAMPLES AND EXPERIMENTS	78
7.1 A re-introduction	78
7.2 A simple example	79
7.2.1 Interpreting graph for “ <i>society civilization culture</i> “	79
7.2.2 Other output files	84
7.2.3 Twisting it a bit.....	87
7.2.4 Contradicting interpretations	89

7.3 Experimenting with corpus data	93
7.4 Experiment 1 - isolated paragraphs.....	99
7.4.1 Paragraph 0	99
7.4.2 Paragraph 1	100
7.4.3 Paragraph 2	101
7.4.4 Paragraph 3	101
7.4.5 Paragraph 4	102
7.4.6 Paragraph 5	102
7.4.7 Paragraph 6	103
7.4.8 Paragraph 7	103
7.4.9 Paragraph 8	104
7.4.10 Conclusion on Experiment 1	105
7.5 Experiment 2 - first and subsequent paragraphs in pairs.	107
7.5.1 Paragraphs 0 and 1 as one.....	107
7.5.2 Paragraphs 0 and 2 as one.....	109
7.5.3 Paragraphs 0 and 3 as one.....	111
7.5.4 Paragraphs 0 and 4 as one.....	112
7.5.5 Paragraphs 0 and 5 as one.....	113
7.5.6 Paragraphs 0 and 6 as one.....	114
7.5.7 Paragraphs 0 and 7 as one.....	114
7.5.8 Paragraphs 0 and 8 as one.....	116
7.5.9 Conclusions of Experiment 2.....	117
7.6 Experiment 3 - paragraphs 0-5 as one - 6, 7 and 8 isolated.	118
7.6.1 New lexemes resulting from experiment 3	118
7.6.2 Lexemes that were removed from contextual relation because of the concatenation of paragraphs.	119
7.6.3 Lexemes that persist in the interpretation of the combined sequences.	119
7.6.4 Conclusions of Experiment 3.....	121
7.7 Experiment 4 - all 9 paragraphs in one sequence.	122
7.7.1 New lexemes in the concatenation of all sequences.....	122
7.7.2 Lexemes that are reinterpreted to fit the context.	125
7.7.3 Conclusions of Experiment 4.....	126
7.8 Conclusions on Experiments.....	126
8 CONCLUDING REMARKS AND FURTHER DEVELOPMENT..	132
8.1 Related research.....	132
8.1.1 The Generative Lexicon)	132
8.1.2 Ontoquery	136
8.2 Evaluation of the project with regard to original goals	137
8.2.1 Sequences of nouns instead of full text.	137

8.2.2 Restricted lexical relationships	138
8.2.3 The scoring scheme.	139
8.2.4 Tracking contextual shifts.....	140
8.3 Conclusion	141
REFERENCES	144
APPENDICES.....	146
A-0 Project description.....	148
A-1 Experiment 1 graphs	150
A-2 Experiment 2 graphs	158
A-3 Experiment 3 graph.....	166
A-4 Experiment 4 graph and glossary	167
A-5 Master results from experiments	182
Table A-5.1 Master scoring table all experiments.....	182
Table A-5.2 Experiment 1 - isolated paragraphs	183
Table A-5.3 Experiment 2 - first and second paragraph.....	185
Table A-5.4 Experiment 2 - first and third paragraph	186
Table A-5.5 Experiment 2 - first and fourth paragraph.....	187
Table A-5.6 Experiment 2 - first and fifth paragraph.....	188
Table A-5.7 Experiment 2 - first and sixth paragraph.....	189
Table A-5.8 Experiment 2 - first and seventh paragraph.....	190
Table A-5.8 Experiment 2 - first and eighth paragraph.....	191
Table A-5.9 Experiment 2 - first and ninth paragraph.....	192
Table A-5.10 Experiment 3 - first six paragraphs	193
Table A-5.11 Experiment 4 - all paragraphs	194

1 Introduction

The present paper is a master's thesis in computer science. I am a bachelor of computer science with a minor degree in linguistics from the University of Copenhagen.

The topic of this work is related to both computer science and linguistics and can be seen as an introduction to the challenges of natural language processing in general and formal semantics in particular.

The main interest of the work described in the present paper is to devise a method for automatically “**skimming**” an informative text in order to arrive at an estimation of the conceptual focus of the text, i.e.: the **semantic context** or more informally just the **context** of the text. I chose the term “skimming” to describe the process because it implies a superficial approach to the content of the text without actually reading and understanding each and every word or paragraph of the text.¹

Such a skimming method would be very useful both in deciding the relevance of the text with regard to some specific topic for instance in search engines or information mining/extraction applications (see for example <http://www.dcs.shef.ac.uk/research/groups/nlp/extraction>), and also as a referential framework to help the in-depth studying of the text either by a human reader or by a computer system.

1.1 Basic Idea

A serious problem in formal semantics of natural language is **lexical ambiguity**, where one word can have several different meanings. As a complement to statistic methods, I propose what could be called a **relational method** to remedy the problem of lexical ambiguity.

The main inspiration for the proposal is the basic intuition that any given text represents a particular **semantic context**. My understanding of such a context is that it involves a particular set of **semantically related concepts** and also particular preferences on just what words to refer to these concepts.

This intuition suggests a **heuristic** for lexical disambiguation that involves a notion of **semantic coherence** based on conceptual relationships:

¹ In fact, the term “skimming” refers to a process performed by a human reader. The process consists of reading the first and last paragraph and perhaps also other compositional clues like the index of an article, in order to capture the main ideas of the text. The process I want to implement here is closer to what is usually called “scanning” when performed by human readers. It involves recognising particular keywords and topics throughout the text in order to narrow in on the content of the text. The term “scanning” does however also have a specialised meaning within Computer Science and therefore I chose the word “skimming” instead.

The best interpretation of an ambiguous word is that which relates semantically the best with the context of its occurrence.

When regarding larger portions of text, the basic idea expands to:

Among alternative interpretations of a given portion of text, the one(s) that involve the most coherent semantic context(s) is better.

In these terms, the development of an experimental system that attempts to establish the semantic context of a natural language text based on the semantic relationships between concepts, is the topic of this paper.

1.2 Processing natural language

Many computer science projects have already concerned themselves with problems pertaining to **natural language processing (NLP)**. Indeed it can be considered as a prominent member of the artificial intelligence family; see for example (Nilsson, Nils J. 1998), (Bratko, Ivan 1990). Also many theories in computational linguistics have of course investigated language and the possible automatic processing of language e.g., (Dalrymple, Kaplan, Maxwell and Zaene 1995), (Sag and Wasow 1999). This kind of work obviously involves many different levels of analysis and for clarity I will point out some of the most important here. Note that this is the set of analytic levels traditionally found in the literature and is what linguists call the **Grammar** of language - using a very specialised meaning of that word; see for example (O'Grady, Dobrovsky, and Katamba 1997).

- **Phonetic Analysis:** is in popular terms the transformation of wave signals into recognizable speech sounds. It also studies how speech organs are configured to produce all the different speech sounds.
- **Phonological Analysis:** is the parsing of speech sounds into phonemes or language components.
- **Morphological Analysis:** is the parsing of phonemes into words or lexemes.
- **Syntactical Analysis:** Is the process of ordering lexemes into larger syntactical structures - sentences, paragraphs, chapters, books and so on. In linguistics and semiotics all of these structures can be seen as (composite) units, sometimes referred to as linguistic **signs**, referring to the notion of the word originating from Ferdinand de Saussure (Saussure 1915, 1974). These units consist of the expressive part on one side and the meaning it carries on the other side.
- **Semantic Analysis:** is responsible for the task of assigning the proper meaning to each expression. What is the meaning of this or that expression, sentence, paragraph, chapter or entire text?
- Intertwining all of the above, a pragmatic analysis ideally should also take place. Pragmatics is the study of situations and circumstances of the language using process (the recorded conversation or written text) as well as the relationships between the participants of the conversation.

Taking into consideration, for example, the author contra the reader of a text, and their respective supposed and actual purposes of participating in the exchange and all other aspects of human existence that could be assumed to affect the process. Pragmatics has only quite recently been accepted as a separate field of study in its own right, within general linguistics; see (Verschueren 1999).

Most approaches to natural language analysis and processing agree to apply phonetics and phonology in one pass to the recorded or perceived sound input, then goes on analysing the result morphologically. Here systems and theories tend to separate. Some apply first syntax and then semantics, each as separate functions, while others strive to establish a link between syntax and semantics. Both syntax and semantics must be recursive (Language allows for infinitely many expressions to be formed and combined, each having a meaning. However complex and unsurpassed the human brain it is not infinitely large, which would ultimately be required for storing infinitely many expression-meaning pairs). This also means that the meaning of sentences cannot be listed in a lexicon like the meanings of words can. The meanings of sentences must be realised as a function of the meanings of the words that go in them, and the way they combine. The notion, that a complex unit is composed as a function of its constituents is referred to as the **principle of compositionality**. If the principle of compositionality is accepted, researchers should look for compositional rules for forming meanings of sentences from meanings of words. In particular the possible constituents should be investigated thoroughly for all compositional information they might hold. (As a corollary, the alternative to accept language as compositional is to say that the meaning of the sentence *cannot* be found solely by investigating the words it holds or the way they combine. For more on this, see (Swart 1998), (Saeed 1997).

With regard to the pragmatic analysis, most applications of computational linguistics have more or less avoided or ignored pragmatic issues and mechanisms, This is because in most cases, the pragmatics can be locked or hardwired in a system designed for functioning in very specific situations solving predictable classes of tasks. Furthermore it is more or less an established fact (one among several that makes formal pragmatics a very complex and intricate matter indeed) that pragmatic mechanisms work on all levels of the above list and that it involves some very intangible measures. The emotional state of one participant, for example, as well as his or her intentions, either real or presumed, can certainly put its mark on both the tone of voice, choice of words, sentence type and so on. All of this has an effect on what is perceived or understood by other participants of the language exchange. In short, if pragmatic issues are at all incorporated in a **NLP** system, it is so indirectly and usually locked into position for the specific task of the system.

To the best of my knowledge, the current state of the art in computational linguistics has made substantial progress in most of the

analytical levels up to and including the basic parts of the semantic analysis, namely the lexical meaning of lexemes, the meanings that can be retrieved by dictionaries and rules of compositionality. This includes resolution of pronouns (resolution of pronouns - or rather *anaphors* - is at the heart of *discourse analysis*, and is by no means a simple task), though some aspects may still be unsolved. The field is currently working at the hard parts of semantic, and to some degree, pragmatic aspects of language understanding.

Automatic translation systems do exist, that are able to produce useable translations of texts between languages - still only in very specific contexts however. PATRANS, for example, is one such system developed by Center for Sprogteknologi in Copenhagen, Denmark. It translates patent applications in the field of biochemistry and physics between, for instance, English and German). One of the things still missing is the general applicability of NLP-systems.

Very few systems exist that are able to analyse data, other than what is restricted to very limited fields of knowledge, where ambiguity is minimal. In order to remedy this, ways have to be devised for automatically recognising the **context** of novel/unseen data (as opposed to data that is pre-constructed or collected for the purpose). Resolving ambiguity of natural language requires access to contextual world knowledge, including, but almost certainly not restricted to ontological knowledge about how concepts are related to each other. The theory of concepts in the world is sometimes called **Ontology**, and reaches far back in philosophical history. Ontologies are also at the heart of the current OntoQuery project; see for example (Jensen and Skadhauge 2001) and (Erdman Thomsen et al 2001).

I adopt in this thesis the **principle of compositionality** as the underlying semantic theory, and attempt to find keys to disambiguation, not in the constituents themselves, but in between them, - in the way they are semantically and **ontologically** related to each other. I attempt to extract and represent parts of the **semantic context** of textual natural language, based on the **lexical relationships** of its **constituents**, and discuss how to compare and exploit this representation.

To make it clear when I refer to concepts and when I refer to word forms, I shall adopt the convention to use SMALL CAPITALS for concepts, and *italics* for word forms and examples. Take the polysemous word form *hammer* for instance. It may refer to a tool used by carpenters for driving nails into wood and, as well, to the ball-on-a-chain object thrown by certain athletes in athletic contests. These two concepts could for ease of reference be called CARPENTERS_HAMMER and SPORTSMANS_HAMMER respectively. Consider then the following examples constructed for the purpose:

- a) "There is nothing that John cannot fix with a *hammer* and some nails."
- b) "John won gold in throwing the *hammer* and silver in javelin."
- c) "John threw the *hammer* through the tool shed window."

d) “John threw the *hammer* through the window.”

In a) the semantic context highlights the CARPENTERS_HAMMER interpretation of *hammer*. In fact the SPORTSMANS_HAMMER possibility is completely absent in this context.

SPORTSMANS_HAMMER is, however, the highlighted one in b), where it would be at best anomalous if the intended meaning of hammer was an object similar to the one referred to in a).

In c) *hammer* is likely to be interpreted as CARPENTERS_HAMMER. This is because this kind of hammer is a tool, and tools are likely to be found in a tool shed, where they could possibly be thrown about for one reason or another. Note, however, that the selective preference at play here in c) is of lesser strength, than is the case in a) and b).

The example in d) is a completely neutral context in the sense, that there is nothing here that prevents one interpretation in particular. Both are in essence possible. Only the status of CARPENTERS_HAMMER as the preferred reading of *hammer* (sometimes also called the *prototypical reading*) primes this interpretation and thus causes it to come to mind more readily. As a corollary, one might argue that the only reason why an alternative reading remains possible is that THROWING is the designated purpose of SPORTSMANS_HAMMER.

I will concern myself with *semantic context* in the rest of this paper. The effects it can have on the interpretation will be illustrated in due time. The syntactic context will not be treated in any depth in this paper.

A good way to think about the context to be treated is as a filter and selector for specific interpretations of ambiguous word forms. It can be seen to involve particular concepts and preference on actual word forms to realize these concepts.

It is my intention to implement these ideas in a prototype-system that accepts a NL-text and produce a series of possible contexts for that text, using the lexical relationships between the possible meanings of words in the text. I will use the prototype for various experiments with real world data and discuss results and perspectives.

The focus of the project is on the meaning of words – and how meanings of words relate to each other in different contexts. The meaning of words - as part of the meaning of entire utterances - falls within the scope of *lexical semantics*.

1.3 Words in context

A Word in isolation has meaning, it hints at and restricts its context.

A naïve approach to lexical semantics could dictate that reading a word in isolation yields no hints as to its *context* (I will talk about context in more detail later in this paper, for now the semantic context or the topic of the conversation in which the words occurs, is what I mean when using this

term), other than the possible meanings of the word itself, since there is no other context present.

This is how it seems - or is it? Well, on closer inspection such an isolated word cannot mean just anything. First of all it does of course only have so many possible lexical meanings. Secondly, one or a few of these possible meanings seem more likely to people trying to understand the message conveyed. They are the *preferred meanings* of the word form. Thirdly, binding the word to any of its meanings has implications on what other concepts make sense in that context.

Consider the word *society*. The preferred meaning is the public society of a community or state, as in: *society has ways of dealing with criminals*. It can however also refer to a small exclusive group or club of likeminded individuals, and maybe more as well. Out of context it is hard to say which meaning to assign to the word *society*. However, if *society* occurs in a text where also *civilization* occurs, it seems to be a reasonably good guess that the political reading of *society* makes the most sense. This is because this meaning of *society* is closely related to one of the meanings of *civilization*, a semantically very restricted word that actually has very few alternative meanings. The relation seems much closer than what relates other meanings of these two words. Together the words make their individual ambiguity much easier to resolve.

Once we have the related meanings of *society* and *civilization*, we immediately have an idea of what to expect from other words in the text to follow. So when *culture* occurs we effortlessly choose the meaning of *culture* that fits with *civilization* and *society*, even though *culture* has several alternative meanings.

This goes to say, that if it is possible to deduce the context of a text (even if only partially), by looking at the words that occurs in the text and compare their alternative meanings, then we can use this emerging picture of what the text is about to: a) further disambiguate the text and b) refine the picture accordingly. The picture will be vague and fuzzy at the start and get more and more clear and restricted as more and more words are disambiguated. Complete understanding of an arbitrary informative text is of course the ultimate dream for the system and not a realistic one at present. Never the less it is the inspiration for the project.

1.4 Aim of the project.

The thesis will attempt to elaborate on this representational idea and attempt to apply lexical relations like the above to “skim” natural language data to extract contextual information from it. I will also discuss how to reason with the context and how it may be possible to acknowledge contextual shifts in longer texts. I will develop and document a prototype program in PROLOG that implements these ideas. It is important to the credibility of the results, that actual data from real world textual language is examined, rather than just a few controlled examples. To keep in line with these guidelines, I will need a suitable corpus and a set of external tools to do the pre-processing. I will also need a **machine-readable dictionary** (MRD) that has the relational information mentioned available. Both the

choice of corpus and the external tools used will respectively be treated in later chapters of this document.

The result should be twofold. Firstly, an experimental representation format for the context of natural language text. Secondly, using actual experiments with the prototype program as a starting point, I will discuss if it seems feasible to automatically recognise a language fragment context - and possibly track changes in the context as the text progresses. I will propose several points of interest in this respect:

- It seems that the ontological knowledge of what words “go” in what contexts can be explained through the lexical relationships of the word meanings. Maybe this notion is part of the key to efficient resolution of certain lexical ambiguities.
- It is likely that the context of one particular text comprises several distinct contexts. It is important to arrive at an understanding of context that allow for contexts to complement each other, while at the same time be able to decide when two contexts are in fact distinct.

The rest of this paper comprises seven more chapters. In chapter 2, I will discuss why I think context representation and recognition is important and what applications would benefit from a deeper understanding of context. In chapter 3, I will describe the necessary theoretical background of lexical semantics in detail. I will also introduce some pragmatic observation that has influence on the project. Chapter 4 describes the MRD that I chose for the experiment and also the external pre-processing tools I adopted and adapted for the purpose. Chapter 5 presents the corpus of informative natural language text I decided to use. Chapter 6 documents the development of the prototype programming. In Chapter 7 I will relate what experiments I completed with the final prototype and discuss the consequences and implications of the results of the experiments. Finally, in Chapter 8, I will present a concluding chapter, relating to what extent the results of the work measures up to my expectations, as well as relate what further work is conceivable and make references to similar work of other people.

2 Natural language processing (NLP)

The topic of this paper is a small yet, I think, important corner of the much larger field of natural language understanding. In the following paragraph, I will make a short introduction to the motivations of research in natural language processing. After this general introduction, I will focus in on the problems and challenges of context representation.

Among the disciplines of artificial intelligence, natural language processing is a prominent member for several reasons. I think that James Allen (Allen 1994) presents a very good and to-the-point formulation of the scientific and applicative purposes of studying natural language understanding. The following excerpt is from the introductory chapter of his book.

“The scientific motivation is to obtain a better understanding of how language works. It recognizes that any of the other traditional disciplines (Linguistics, psycholinguistics, language philosophy and others .ed), does not have the tools to completely address the problem of how language comprehension and production work. Even if you combine all the approaches, a comprehensive theory would be too complex to be studied using traditional methods. But we may be able to realize such complex theories as computer programs and then test them by observing how well they perform. By seeing where they fail, we can incrementally improve them. Computational models may provide very specific predictions about human behaviour that can then be explored by the psycholinguist. By continuing in this process, we may eventually acquire a deep understanding how human language processing occurs. To realize such a dream will take the combined efforts of linguists, psycholinguists, philosophers, and computer scientists. This common goal has motivated a new area of interdisciplinary research often called cognitive science.

The practical, or technological, motivation is that natural language processing capabilities would revolutionize the way computers are used. Since most of human knowledge is recorded in linguistic form, computers that could understand natural language could access all this information. In addition, natural language interfaces to computers would allow complex systems to be accessible to everyone. Such systems would be considerably more flexible and intelligent than is possible with current computer technology. For technological purposes it does not matter if the model used reflects the way humans process language. It only matters that it works.”

One of the very hard parts of language understanding is the semantic task of assigning meaning to words and seeing how these combine to give meaning to sentences, and how sentences relate their meanings to each

other. The fact that word meaning itself can be very difficult to pinpoint is easily verifiable, even for humans. World knowledge has an important part to play, a part that lies close to the core of the problems associated with lexical ambiguity. That is to say, that to know when to use one word instead of another, and when one word means one thing and when it means something entirely different, it is necessary to have experienced the different situations or at least have them described in detail. Since we cannot expect computers to experience anything what so ever (not in the usual meaning of the word anyway), we will have to make do with describing as much as possible of what is necessary to them, and grab on to all the hints that might be present in language itself. In this respect the notion of “semantic context” is relevant, as possibly supplying most of these hints as to where in the world to place the data at hand.

2.2 Context

When the child, in doubt of when a certain word means one thing and when another, ask its parent or teacher for clarification, the answer received could be that the context will show what the meaning of the word must be. An understanding of the context is central to pinning down the meaning of ambiguous lexemes.

So what is “context”?

The term “context” is, if any, truly a polysemous one. First there is the syntactic context of discourse analysis. This term refers to those words that co-occur with a particular word. Conversely the term “context” is often used to refer to the concepts that these words realize, we could call this the semantic context. The understanding of context that I use in this paper is a reference to those lexemes that co-occur in a particular semantic domain. For the purpose of this paper I propose the following definitions.

Context (def.) A Context is an unambiguous selection of lexemes that either

- a) are all semantically related in one coherent component.
- b) is the combination of two or more distinct contexts.

Lexeme (def.) A Lexeme is a pair consisting of a concept and a specific orthographic word (and inflections) that can realize that concept. That two lexemes are semantically related means that either their respective concepts are related, or that the words themselves are related.

On one side, this definition allows us to relate arbitrary descriptive/informative texts to the context(s) they describe. On the other side we can see a similar relation between the meaning(s) of an arbitrary word and the possible contexts they are related to.

Going in one direction it seems possible to regard the meanings of the words of a text and get the possible contexts that the text might describe.

Having a general idea of the context described, it seems possible to go the other way as well, and restrict the ambiguity of words in any text describing that context.

2.3 Ambiguity

Lexical ambiguity is among the harder problems of natural language understanding. In this respect there are two main categories of lexical ambiguity:

Homonyms are words with several unrelated senses; the word *lap*, for instance, can refer to both a full circle in racing sports, and to the upper part of the legs of a sitting person.

Polysemes are words with several distinct but related senses. For example the word *society* in the previous chapter has at least two different but still somewhat similar readings.

In the sentence, *We have a powerful society*, it is impossible, even for a human reader, to decide which of the two meanings to assign to the noun *society*, since both seem to make sense. (The only clue here is the preferred reading of the word)

In *John just finished a lap*, a human reader will have no difficulties in deciding that, John must be involved in some circular motion, since the leg-reading of *lap* is for children to sit on and clumsy people to spill sauce in. It cannot be *finished* in any way (except perhaps from cleaning, after the sauce is spilled. Note that given circumstances that are sufficiently absurd, all sorts of readings are possible though unlikely).

Enabling a computer program to make these distinctions is not trivial at all. An entire network of information about the world and intricacies of human endeavour and behaviour is necessary to even begin to solve this complex task. Mostly the information is available to the human reader in the context, though unintelligible phrasings, even for a human reader, occur all the time. Some of this information may be available to the computer program as well, if the context can be properly represented.

It can be argued that any successful NLP-system must, at the very least, pay attention to what contextual information is present in the data, or it will lose out on essential keys to solving the semantic task of reliably assigning meaning to its input. There is ample motivation to experiment with how to define and represent the context of natural language, and how to extract as much of it as possible.

2.4 Applications of skimming

Besides the important benefit to disambiguation in a larger NLP system, a working skimming system can also be seen to have applications of its own.

Consider, for instance, the ability to recognise the proper context of emails in your private mailbox. To a large degree, avoiding spam mail is extremely hard to do by searching for specific keywords in the mail. The sender can all too easily rephrase the wording of the spam mail to contain nothing but “innocent looking” words and thus evades the schemes for blocking it. If however a system could be able to recognise relationships between words and compare them to a list of suspicious relationships, it would be much harder to bypass. The ability to apply a familiar word in a new and novel context is one of the hallmarks of natural language, and in order to recognise this kind of creative word use, employing transferred meanings and metaphors, it is necessary to know the different contexts of word usage.

Another obvious application is as a part of a search engine to the World Wide Web, possibly enabling these to cut down on the number of links that are *irrelevant* to the user. Again, a set of search words could be combined with a restricting set of relationships between them. The system would then compare this context to the contexts of the candidate links, to decide possible relevance to the user. This comes close to the goals formulated in the OntoQuery project see (Jensen and Skadhauge 2001), (Erdman Thomsen et al 2001).

These are but a few examples of how the proper function of the ideas treated in this thesis could be used. I have no doubt, however, that there are many possible applications, and that it will also take time to refine the natural language processing capabilities of computers to a degree where these goals are feasible. Examining how the semantic context can be represented, and possibly recognised, does however seem to be a small but important step in that direction.

3 Theoretical analysis

In this chapter I will present the theoretical background necessary to write a computer program that reads actual natural language text and reasons with the general context of that text. The ideas and notions described in this chapter are what first led me in the direction of this project and indeed form the theoretical basis for the proper functioning of the prototype and any systems developed from it. Therefore they deserve a considerably detailed treatment and this I try to respect in this chapter.

First I repeat the typographical conventions, and then I introduce some technical terms that may not have been treated elsewhere and after that I will define the scope of the task at hand.

3.1 Conventions and decisions

3.1.1 Typography

I have introduced the conventions that the “graphical words” occurring in data and examples are represented in *italics*, while the meanings are in SMALL CAPITALS. This way, italics represent the actual, possibly, ambiguous occurrences of words in data, while the small capitals will represent concepts. The actual wording used for describing concepts is not in any way fixed but chosen in each case to explain the meaning as clearly as possible. Some terms may be in bold face small capitals like NUMBER and PLURAL. These terms are for paradigmatic attributes and their values.

3.1.2 Technical terms

I am going to touch upon subjects that are not really within the topic to explain the choices I make. To avoid confusion and ensure the clear understanding of my examples without making the digressions too far reaching, I am summing the definitions up here. The terms “Lexeme”, “word” and “expression” are in this respect almost parallel to the terms “concept”, “meaning” and “content”. The first three refer to the carrier of some content, and the others refer to the content itself, the senses of the expressions.

Lexemes, words and expressions.

These terms all refer to units that can have meaning. I will use the term “expression” to refer to any such carrier of meaning in general. I will use the term “word” to refer to a particular form, having possibly several meanings. I will use the term “lexeme” to refer to the paradigm of a canonical form and all its inflected forms with one particular meaning. A lexeme is said to “realize” the concept associated with it.

So if we consider these two words and regard them each as unambiguous:

- a) *pebble* (SMALL_ROUND_STONE)
- b) *pebbles* (SMALL_ROUND_STONE + PLURAL)

I will say that they are different forms of the same lexeme. I will also say that the uninflected form in a) is the canonical or orthographic form of the lexeme.

- c) *wood* (COLLECTION_OF_TREES)
- d) *wood* (ORGANIC_MATERIAL)

In c) and d) we have two instances of an ambiguous word, and I will say they belong to different lexemes because they realize different concepts. The same is the case for e) and f) below.

- e) *sail* (TO_MOVE_IN_A_BOAT_OR_SHIP_ON_WATER)
- f) *sail* (SHEET_OF_FABRIC_USED_FOR_PROPULSION)

I must point out that this definition of the term lexeme is closely related to the particular lexicon I will employ for the experimental system. There exists a host of slightly different definitions of the term in the literature and it is not the task of this paper to suggest or discuss which is better or more correct. The lexicon I will use adopts this definition and so must I. The actual choice is not however critical in any way to the core ideas of the thesis, as long as the different meanings of a word can be distinguished. The definition used, however, decides when to list words like *wood* in two different entries, because it represents different lexemes, and when to have only one entry representing several forms like *pebble* and *pebbles*, because they belong to the same lexeme.

Concepts, meanings and content.

These terms all refer to the semantic side of the line between an expression and what it expresses. The meaning of a sentence may involve several concepts, while the meaning of a lexeme is usually one concept. So when discussing a lexeme and its semantic content I will speak of the “meaning” of the lexeme. When the actual form is not important, I will talk about “concepts” instead. Abstracting further away I will use the term “content” to refer to the semantic side of the above mentioned line in general.

Inflection.

When *boy* and *boys* are said to be different forms of the same lexeme it is because of inflection. The –s suffix marks the PLURAL aspect in *boys*, absent in *boy*. Inflection is a regular and predictable process. For English nouns only the PLURAL/SINGULAR distinction and the possessive –s exists. For English adjectives there is the distinction between degree POSITIVE, COMPARATIVE and SUPERLATIVE. For verbs, more inflections are possible, for instance, TENSE, ASPECT, NUMBER, PERSON. In short, inflection concerns syntactical variations of the same lexeme. The task of deciphering the effects of inflection is a morphological one and will not be treated in any further depth in this work.

Derivation.

Derivation transforms one lexeme into a new one. A good example is the noun *beauty*. When used with the derivative suffix “-ful” it becomes the adjective *beautiful*, and with the derivative suffix “-fy” it becomes the verb *beautify*. *Beauty*, *beautiful* and *beautify* are all different lexemes.

Since the meaning of either *beautiful* and *beautify* can be said to include the meaning of *beauty*, *beauty* can be considered the more basic of the three.

Beautiful and *beautify* are said to be derived from - or derivations of - *beauty*. Like this there is a whole range of derivational rules, each with its own suffix and each with their own purpose of incorporating one concept in another. For example a derived form like *beautiful* can easily undergo further derivation, namely from adjective to adverb using the adverbial suffix “-(l)ly” to result as a new lexeme : *beautifully*. Because derivation can change a word’s class, the resulting word must per definition be a different lexeme. One lexeme can only belong to one class (and sense in my definition). Like inflection, resolution of derivation is a morphological task that will not be described in particular in this work.

3.1.3 A word on illustrations, figures and examples

This chapter incorporates numerous illustrations, figures and examples. Some of them are of my own device, some are used from various sources in the literature, and some again I have adapted from such sources to suit my needs. Whenever other people should be credited they will be in the figure text or in close association with the example. If no such reference is present, the work is my own.

3.2 Definitions and restrictions

3.2.1 Definition of context

To continue the simplicity of the ideas developed in the previous chapters, a context will for the time being be seen to simply comprise:

- A set of particular concepts.
- A set of particular words realizing these concepts.

This is obviously not the whole truth about contexts, but it will serve for now.

3.2.2 Restrictions on data

To begin with some consideration has to be taken to what kind of language to analyse. It is not realistic to apply an “understanding” algorithm to just any kind of language, since not all linguistic expressions are intended to be understood in the usual sense of *understand*. Poetry and short stories for instance can arguably be said to be for experiencing rather than understanding and we can’t - at least not in any foreseeable future - expect computers to be able to experience and appreciate the contents of any such artistic expression. With the practical applications in mind, it will therefore be helpful to restrict the linguistic data suitable for automatic processing to

be *informative*, in that it must contain factual information that is intended to be immediately understood.

3.2.3 Restrictions on analytical levels

Once the data has been decided upon, comes the time to decide how to treat it. As pointed out in the introduction of this paper, processing natural language requires treatment of all the analytic levels of *phonetics, phonology, morphology, syntax, semantics and pragmatics*. Because I want to investigate written language, the sound related levels of analysis are eliminated. The remaining levels of analysis still pose a set of tasks that are much too complex to be treated within the scope of this paper. My chosen focus is on mechanisms particular to the semantics and pragmatics, but I still want to use actual language for data. In order to get to the semantic and pragmatic parts without having to deal with morphology and syntax, I will have to employ external applications for analysing morphology and syntax.

3.2.4 Restrictions on concepts

The mechanisms I want to explore are the relationships between concepts, but concepts can arguably be of many kinds. It could be said that `RUNNING` is a concept as well as `RED` and even sentential concepts can be perceived, `PLAYING_THE_VIOLIN` for example. This comes close to the notion of Ontology – the study of the categorical structure of reality (Nilsson, Nils J. 1998). It leads too far to present a complete discussion of ontologies and every kind of concept. Prototypical concepts, however, are nominal ones like `BIRD`, `HAMMER` and `SOCIETY`. It is also in the domain of nominal concepts that semantic relations are most easily studied. This probably has to do with the way humans perceive the world around them and make “things” in the world “fit” into the conceptual system of their mind. Physical objects and entities are the kind of phenomena that most readily lend themselves to this kind of categorization. Physical objects and entities are in the domain of nouns.

For reasons of simplicity and coherence I shall restrict the experimental prototype to treat only the nouns occurring in the data. Where this choice has consequences for the functioning of the prototype program I will discuss those problems, as they occur.

3.2.5 Restrictions on accuracy

To reason with the meanings of words the prototype must have access to a suitable dictionary or lexicon formulated in a way that can be accessed as easily as possible by a computer program. This **machine-readable dictionary (MRD)** will in this case be static in the sense that its explanations will not change or evolve through experience. In this way it differs fundamentally from the mental lexicon of a human language user, that is in ongoing change. The MRD will, in this case, be like the paper dictionary on the shelf where unknown but established word meanings and their relationships can be looked up and used for clarification. Due to the static nature of the MRD, the system must necessarily make do with the

information present at the time of publishing of the lexicon. That means that idiosyncrasies of era and society are in a sense inherited from the MRD. This way the result cannot be in complete accordance with all aspects of the language because languages are in endless change and evolution. This may seem as if it is a serious flaw, but it is not necessarily so. I can simply adopt the view that the information in the MRD actually **is** accurate with respect to the language at hand.

3.2.6 English as data language

The necessity of external tools and the restricted availability of such tools dictate that I choose English/American for the language of the corpus or data set.

3.2.7 The experimental tasks

With these restrictions in place the overall task can be identified as follows.

- Find a suitable English language corpus of informative text. In the corpus, identify paragraph limits and find all occurrences of nouns within each paragraph. The task on finding a suitable corpus is one for me to complete, while establishing paragraphs boundaries and identifying nouns will be done by external tools.
- For all the nouns of a paragraph find the concepts they may possibly realize, i.e. find their alternative senses. This is a task of the prototype program in interaction with the MRD.
- Find among the senses of each nominal word form the sense(s) that best “fit” with the senses of other nominal word forms in the paragraph. This again is a task for the prototype program, and will draw on the relational information also present in the MRD.

Theoretically the most accurate interpretation of the data is the one consisting of the most coherent and closely related concepts. The concepts found in the way described above will include senses that both fit together well and are possible interpretations of the word forms in the data. The next task for the thesis is therefore to experiment with application of the experimental prototype to different parts of the corpus. The experiments will look for indications of how the context chosen by the prototype relates to the actual context of the data as it would likely be conceived by a human reader. I also call this human-conceived context the *subjective context*, as opposed to objective context of the prototype system.

3.3 Semantic theories and philosophies

The notion of natural language semantics usually refers in general to all of the semantic aspects of natural language. It affects both the generative process of choosing a formulation to convey the intended meaning on the sender(S)'s part and the process of assigning meaning to the perceived utterance on the receiver(R)'s part.

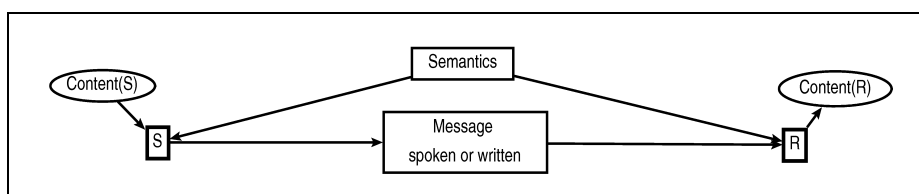


Figure 1 Simple communication diagram

In my simple diagram in Figure 1 is illustrated how S has something to say, so to speak, represented by Content(S). He applies the semantics of the language he wants to use and forms his utterance using words he thinks will best convey that content. The result is the message sent. In this very simple representation that is also what R has to work with. In order to understand the received message R on his side applies the semantics of the language to the message to make sense of it. If the transaction is successful then Content(R) is at least comparable to Content(S). This means that R has not only managed to make sense of the message, but also the right one, namely the one S intended for R to understand. For this to be possible the participants must share the semantic component in sufficient degree. There may be reasons to why this is not always the case however, and I will touch on this in a little more detail when I describe pragmatics in section 3.4.4. The above diagram is of course a grossly oversimplified depiction of a language exchange. It does however introduce several basic relationships and notions necessary to understand the finer points of any semantic theory for natural language; and the picture will be refined in due course.

For the purpose of the prototypical program under development, the S is the author(s) or publisher of the original text used as the experimental corpus and R is the computer program prototype itself. The message to be “understood” is the text of the actual corpus.

Well not quite. The construction of a formal semantic theory for the English language is really, as pointed out earlier, a monumentally complex task and completely outside of the scope of the project. I have stated that I will concern myself with the nominal word forms of the data, and on their internal relations in an attempt to get a superficial understanding of what the text might be about - its context. So the message in the above diagram is comprised of a series of nouns in various inflexions.

The task of assigning meaning to words is the responsibility of lexical semantics. Necessary parts of lexical semantics will follow in the next section.

The other important aspect omitted in the diagram is the situation for the actual language exchange. As hinted at in the introduction of this paper, all kinds of influences can have effect on both the actual message formulated by S, as well as the content understood by R. This paper will not deal in any detail with these abnormal uses of language, however, since in the vast majorities of cases, the sender wants his message to be understood and makes an effort for his choice of words to serve that purpose. To do that

he adopts the conventions of the **Cooperative Principle** as will be described in the end of this section.

3.3.1 Lexical semantics

There are many ways one can approach the lexical semantic task, some formal and some descriptive. I shall not attempt to describe more than the one that I find most accurately relates to my idea of how words make sense. Cruse does, in his very illuminating and well written textbook, *Lexical Semantics* (Cruse 1986), also adopt this approach and refers to (Haas 1962) and (Allerton et al 1979) for further information on the **Contextual approach to lexical semantics**. At the core of this approach it is assumed that:

“ - the semantic properties of a lexical item are fully reflected in appropriate aspects of the relations it contracts with actual and potential contexts.”

Cruse here uses the term “lexical item” and perhaps it is time to be a bit more specific about the object under investigation. Words have the ability to carry information; this is the sign in Ferdinand de Saussure’s terms, (Saussure 1915, 1974), depicted below.

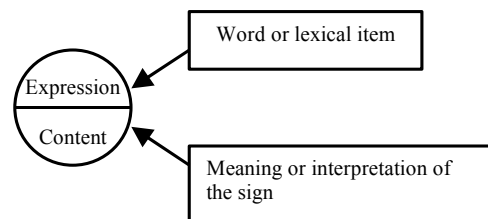


Figure 2 Ferdinand de Saussure's basic "sign".

The distinction between the content and expression sides of the Saussurean sign is important, since recognising and parsing the expression side is necessary to access and reason with information related to the content side. It is not so, that once the expression is properly received, its content is automatically known, not by machines and not by humans either. Knowledge of semantics and pragmatics of the particular language is necessary in order to make any reliable interpretation.

In these terms a text or language utterance in general would be seen as a series of expressions and contents that combine to produce the content of the whole utterance in accordance with the semantics. The following description follows closely that of (Saeed, John. I. 1997) in his treatment of Saussure’s influence on natural language semantics theory as a whole. Each word has meaning on its own, but also connects to other words in the same language like a cell in a network. Figure 3 below depicts how each word refers to actual entities in the world. The semantic links between elements in the vocabulary system is an aspect of their sense. In the figure each circle is a sign, expression–content pair, and the arrows between them symbolise the

semantic relationships that make up their interconnectivity. The figure is a simple one, that does not really do justice to the fact that each sign may have multiple connections, and there is not necessarily a connection between each and every pair of signs. Furthermore the connections may be of both the two-way and one-way variation. The figure does however attempt to make the distinction between the two layers of expression and content, and hint at their relationships.

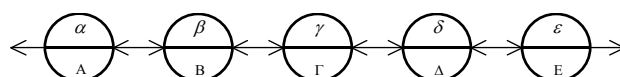


Figure 3 The network of referential signs (adapted from Saeed, John. I. (1997)).

So if we set for instance α to be the word *civilization* then A must be the meaning CIVILIZATION, (incidentally this particular word comes close to being completely unambiguous). If B is then the meaning SOCIETY, then the relationship between CIVILIZATION and SOCIETY (hyponymy) is reflected in the link between the two circles in the above diagram. This connection will be relevant whenever *civilization* and any word realizing the meaning SOCIETY - *society* itself for instance - co-occur, i.e. are present in the same portion of data. Whatever strength lies in the simplicity of this model, it is important to remember that the relationships between expressions and contents are not exclusive. By this I mean that there may be several expressions involving one and the same content, and as well, more than one content may be associated with a given expression.

3.3.2 Synonymy and polysemy

There are two main reasons why the relationships between expressions and contents are not one-to-one. Synonymy and polysemy are most certainly relevant to the task at hand. Synonymy is the semantic relationship between lexemes that have the same meaning. If we accept that two words may have “the same” meaning in particular contexts, it is sometimes helpful to regard the synonymous words as members of the same mathematical set of words that can realize the sense in question. This leads to a new abstraction of the relationship between word and meaning as shown in Figure 4.

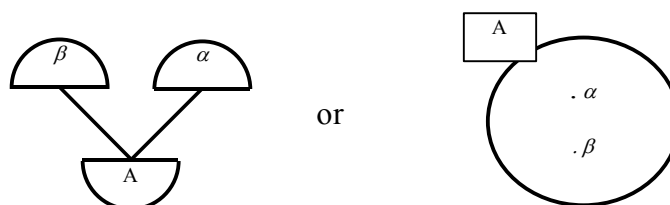


Figure 4: Two ways of representing synonymy. To the left both word forms α and β realizes the concept of A represented by the connection between the words and the sense. To the right the same thing is represented. The fact that α and β have synonymous interpretations as realising A is seen to the right by having them both as members of the synonym set A.

Note, that this is not a redefinition of synonymy. It is merely a slightly different abstraction of the same thing. This representation allows us to distinguish between word form relationships and relationships between senses in an orderly manner. Word form relationships, apart from synonymy, will be marked as arcs between particular set members, and sense relationships will be marked as arcs between particular sets. I shall be using the two representational styles interchangeably as best to illustrate my points.

Resolution of synonymy by hyponymy

True synonyms are in fact occurring very rarely in the lexica of human languages. The two English words *murder* and *killing* have nearly the same meaning when read as nouns, and can in the proper context be used interchangeably with no loss of meaning (The use of *murder* and *killing* in these examples was inspired from a lookup in Roget's Thesaurus, but the actual examples are of my own devise):

- a) *Jack the Ripper committed a series of horrid murders.*
- b) *Jack the Ripper committed a series of horrid killings.*

As such they are synonymous but still, an *accidental killing* is not a *murder*. Similarly an *execution* is obviously a kind of *killing*, and might very well be seen as *murder*.

The point is that the meanings of words may overlap and differ to various degrees, and where circumstances are not concerned with the differing aspects, the words may be used synonymously as in the example above; but where the circumstances are sensitive to the differences in meanings interchanging use will sound odd at best, possibly even misleading. This is represented in *Figure 5* below, where the meaning of the three words are very close to each other and in fact overlapping, but still discretely differing.

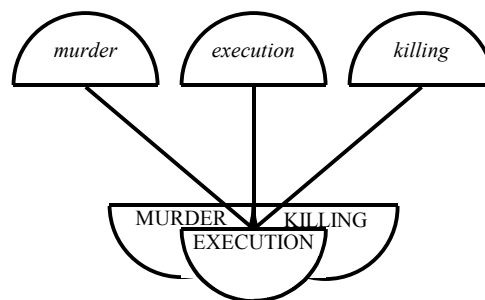


Figure 5: It looks like the words have meaning in common?

This description is still too fuzzy to be completely satisfying though. We need a way to account for their similarity and their differences in a more precise manner. So let's look a bit more closely at the meanings. Upon close

examination, we will see that the meanings themselves are related to each other rather than overlapping. First let's agree on the following description of the three meanings:

- MURDER: the unlawful wilful causing of someone's death.
- EXECUTION: a society's lawful punishment of a criminal offence by death to the offender according to law or codex of the society.
- KILLING: the causing of someone's or something's death.

(The following examples adopt the use of "*" in front of something semantically unacceptable, and "?" in front of something arguably unacceptable. The examples are again my own devise.)

Already a structure emerges, since both of the following:

- a) *A murder is always a killing.*
- b) *An execution is always a killing.*

hold true according to the definition while neither of the ones in c) and d):

- c) ** A killing is always murder.*
- d) ** A killing is always an execution.*

are true. This indicates that both MURDER and EXECUTION are hyponyms of KILLING.

- e) ** A murder is always an execution.*
- f) *? An execution is always a murder.*

The postulate in e) is obviously false, while the truth of the one in f) is a bit questionable. With the readings pointed out above MURDER is unlawful, and EXECUTION is lawful and that would entail that f) is false. Suppose however, the lawful/unlawful part of the descriptions were left out:

- MURDER 2: the wilful causing of someone's death.
- EXECUTION 2: a society's punishment of a criminal offence by death to the offender according to law or codex of the society.

The sentence in f) would hold true and EXECUTION 2 would be a hyponym of MURDER 2. This illustrates how the hyponymy of meanings could help pinpoint their differences. The relationships can be depicted as in *Figure 6*.

This representation operates with two different senses of the two readings of *murder* because they behave differently in their relation to various contexts. As such *murder* and *execution* are both polysemous due to the fact that they can refer to several different senses. As with synonymy it is important that I may show this relationship in two different ways. Again regard senses as sets of their realizing words and observe that the two representations in *Figure 7* illustrate the same thing.

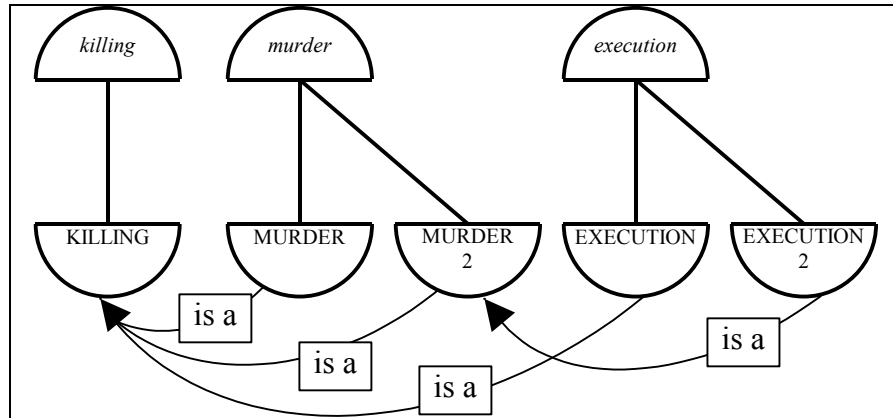


Figure 6: The Semantic relationships between senses reflect the polysemy of the word forms realising them. Note that since the *is_a* relation is transitive, the fact that EXECUTION 2 is a KILLING, follows from EXECUTION 2 is a MURDER 2 and MURDER 2 is a KILLING.

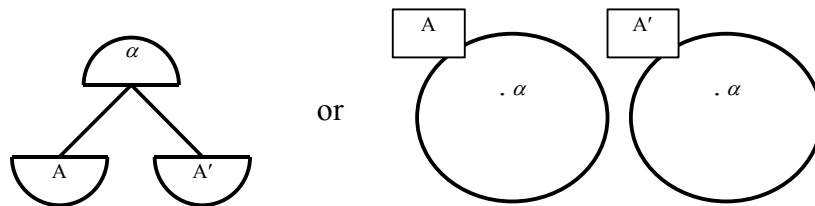


Figure 7: A word form that has several different but still somewhat related readings is said to be a polysemous word form. I show Polysemy in one of two ways. To the left the word form α realises both A and A'. This is represented by the connection to both contents. To the right the same thing is represented in having α as member of both sense sets A and A'.

Polysemy

Polysemy is the other main reason – apart from synonymy – that the mapping between expression and content is rarely a one-to-one affair. Polysemy is defined as the property of a word form that has several distinct but still related meanings for instance “to murder” and “a murder”, where one is a verb and the other is a noun. Obviously these two meanings are different, but quite clearly they are related. Thus the form *murder* is a polysemous one.

Furthermore, it is primarily the senses that are subject to the semantic relationships like the hyponymies of Figure 6– rather than the words. Adhering to the contextual approach to lexical semantics, a comparison of the semantic relationships of the possible interpretations - or senses of a lexical item and the actual context of its occurrence - will ideally identify which reading to associate with the occurrence of a word. Now let’s take a look at the different lexical relations.

3.3.3 Lexical relations

The term “lexical relations” is actually a bit misleading because, as just pointed out, it is the senses that are related in the vast majority of cases, not the word forms, so “semantic relations” or “conceptual relations” might be better choices. However lexical relations are traditionally defined in the literature as ranging over lexemes, that is - over expression-sense pairs, and the use of the term “lexical relations” is so widespread that a redefinition here would not help clarity. I will therefore stick with the definition that lexical relations are relations that range over lexemes, but point out that most of them range over the sense part of the lexemes, while a few others range over the expression part of the lexeme.

To help facilitate the proper understanding of this it seems necessary to make completely clear what I mean by lexeme. Therefore in the following, when I refer to a lexeme I mean a specific word (and its inflections) paired with one particular meaning. So if a word can have three different meanings, the word form will be representative of three different lexemes.

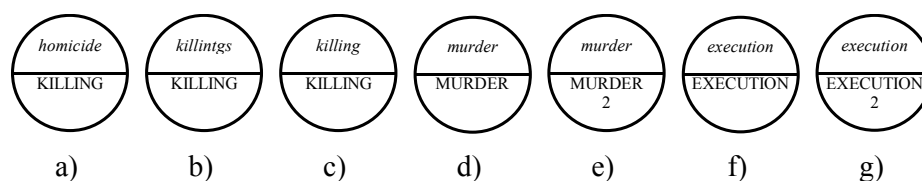


Figure 8: Seven signs, five words, five senses and six lexemes. Note that only *killing* and *killings* belong to the same lexeme, *homicide* and *killing* are synonymous while *murder* and *execution* are both polysemous in this simplified representation. Furthermore, note that the extra aspect plural to *killings* in b) is not represented in this figure. *Killings* in b) still refer to the same KILLING as does *homicide* in a) and *killing* in c). Note that all these forms are nominal, and that verb-readings of *killing* and *murder* was excluded.

In *Figure 8*, are seven instances of a word form being paired with a meaning. The representation is not exhaustive, in that any of the word forms may have other meanings than the ones represented, as well as any of the senses may have other realizing word forms than the ones depicted. The senses are all nominal senses, as to keep in line with the task at hand, though both *killing* and *murder* may be interpreted as verbs as well. b) and c) belong to the same lexeme since they only differ in the predictable inflective aspect of NUMBER. All others are easily seen to be different lexemes. In a) *homicide* is interpreted as a synonymous lexeme to *killing* in c). a) and b) would be synonyms had it not been for the plural aspect of *killings* absent in a).

Arguably it may seem to go a bit far to discuss lexemes, since I am not going to treat morphology in any serious depth. The notion of lexeme is, however, important to understand why *execution*-EXECUTION₂ is a hyponym of *murder*-MURDER₂ whereas *execution*-EXECUTION is not. Having different senses, *murder* is representative of different lexemes, and only one of them is in hyponymous relation to the *murder*-MURDER₂ lexeme, because hyponymy is defined over senses, not word forms. With this in mind it is

now time to take a look at some of the different lexical relations in a little more detail.

Synonymy

The relation of synonymy has already been covered as a reason why the mapping from expression to content is rarely one-to-one. Unlike polysemy, synonymy is important also in the fact that it behaves almost as if representing the identity function of word forms. As such it is one of the few lexical relations that range over word forms rather than senses. This aspect of synonymy is often used in actual language exchange as an explanation tool.

a)

speaker 1: There was *a killing* in the street last night.

speaker 2: What?!

speaker 1: You know - *a homicide*. A guy was shot!

speaker 2: Oh no!

In this example the fact that *killing* and *homicide* can be interpreted synonymously and that *homicide* is considered unambiguous, helps to pin down the meaning of *killing*. It can be seen as if speaker 1 introduces a word form α that can mean both A and A'. If speaker B has trouble choosing between the two and asks for clarification speaker 1 can help by introducing a word form β that means only A.

Antonymy

Antonymy is the relation between a pair of opposites. Obvious examples are

- *Boy and girl,*
- *Man and woman,*
- *Man and beast,*
- *Light and darkness,*

(Note that, even all my examples above are nominal, antonymy operates over many word classes. Consider *easy* vs. *hard*, and *run* vs. *walk*, for instance.)

Like synonymy, antonymy is a relation between lexemes more than between senses alone. Since none of the above pairs allows for one of the words to be substituted with a synonym while retaining the same contrast. It is clear that they are not really strict opposites either, since both *boy* and *girl* refer to human offspring. They are opposites in the only dimension where they are positioned on each extreme, namely gender. That goes for *man* and *woman*, and *bachelor* and *spinster* as well, but as it can be seen below they are alike in all other categorical aspects.

This has led to the notion of “meaning components” as atomic units to meaning. Devising a complete and coherent list of such atomic meaning has

been found to be indeed very hard and I am only mentioning it here as a digression.

To return to the subject, it is obvious that explaining the meaning of a particular lexeme in terms of its antonyms along various dimensions should be a very powerful approach. To say, for instance, that *a boy* is the opposite of both *girl* and *man* does a lot to explain what meaning is intended when using the word *boy*.

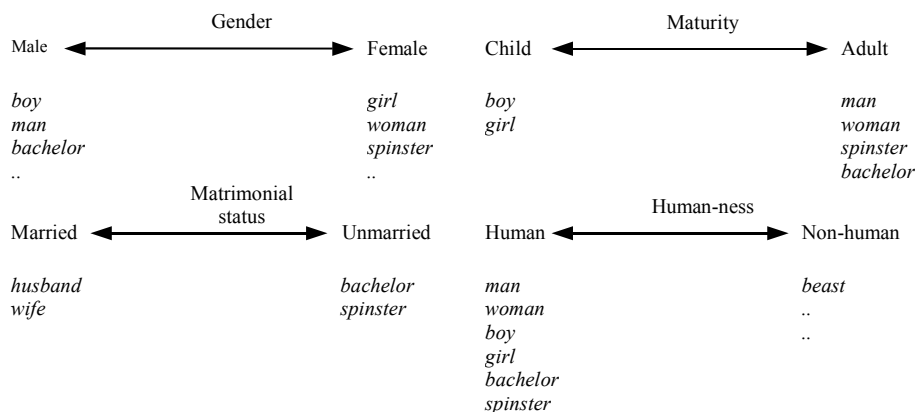


Figure 9: Defining extremes of four different semantic dimensions. Notice how the ordering of the concepts helps describing their differences and interrelations.

Interestingly enough, the dimensions along which antonyms are opposites seems to play a role in the way hyponyms arrange themselves. For instance, the fact that *a boy* is a male human child, also seems like a complete categorical description of the lexeme *boy-BOY*. This is true also if we substitute *lad* for *boy*: *a lad* is a male human child. But *lad* and *girl* are still not antonyms. This places antonymy among the relationships ranging over the lexemes themselves, i.e. over particular combinations of word and sense.

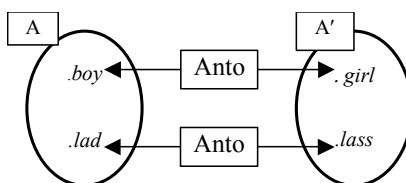


Figure 10: Antonymy, like synonymy, ranges over words with particular meanings (what I call lexemes). Unlike synonyms the connection between the antonyms lexemes must be made explicit by the labelled arc (recall that the connection between synonyms is seen by co-membership of the same set).

Hyponymy

I have already touched on hyponymy as an example of how the co-occurring words (co-occurring meaning occurring in the same portion of data) may help to clarify the proper sense of a polysemous word. Hyponymy can be defined as the inclusion of a particular concept within a more general

concept. It reflects - and was most likely refined originally for the description and categorization of – the natural kinds:

- a) *A Siamese is a cat.*
- b) *A cat is an animal.*
- c) *An animal is a living being.*
- d) *A living being is a physical entity.*

From the example it is easy to see why hyponymy is also called the “is_a” relation. For clarity *Siamese* is said to be the hyponym of *cat* and *cat* is said to be the hypernym or superordinate of *Siamese* in this meaning and from the term hypernym is gained the reverse relation of hyponymy. The inclusion relation is obviously transitive in that:

- if A is a B,
- and B is a C
- then A is a C.

This way hyponymy is alike to “<” for natural numbers in mathematics and hypernymy is like “>”.

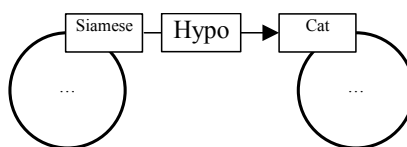


Figure 11: Hyponymy ranges over concepts.

The key here is to agree on which concepts have the lowest possible value. These are called “unique beginners” in the literature and intuitively form a set of very general, semantically almost empty concepts such as entity, phenomenon and so on. The unique beginners have in common that they have no super-ordinates, and that together they have all other concepts as sub-ordinates or hyponyms.

Together the sub-ordinates of for instance Animal, form the set of all kinds of animals, and the sub-ordinates of Cat form the set of all kinds of cats. Such a partially ordered set of concepts directly satisfying a common prerequisite is called a taxonomy. (So SIAMESE is a taxonomic member of CAT, since it is an immediate subordinate to this concept. This does NOT hold between for instance SIAMESE and ANIMAL, because even though SIAMESE is sub-ordinate to ANIMAL, it is so indirectly; namely via CAT.)

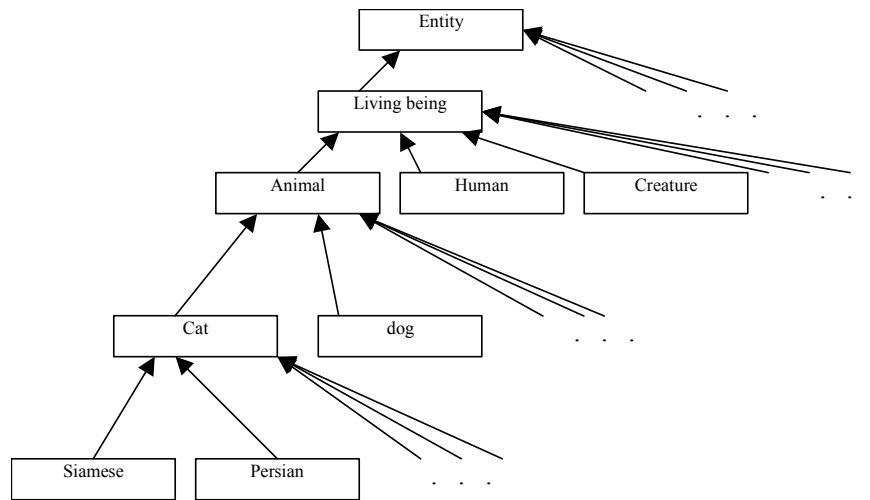


Figure 12: Taxonomies of Natural kinds and Hyponymy.

Meronymy

This family of relationships are in fact special kinds of taxonomies. It refers to the relation of a complex to its constituting parts. Several slightly different kinds of meronymy are identified in the literature.

- Part-to-whole is what relates the concepts of paw, whisker ,leg, tail, ear ... to cat for instance.
- Part-to-substance is what relates drop to water
- Part-to-mass is what relates grain to sand
- Member-to-group is what relates, well, member to group
- And several others as well.

Where I deem it useful I will use the term meronymy to refer to any instance of this kind of relationship. Meronymy ranges over concepts, that is, over the sense part of the lexemes in my definition.

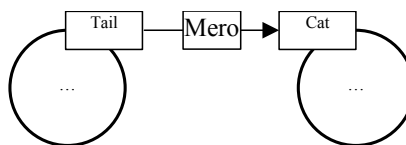


Figure 13: Meronymy also ranges over concepts rather than lexemes.

3.3.4 Pragmatics - the Cooperative Principle

I have described some of the different lexical relations, and showed that they play a role in the minds of human beings. Whether a particular person recognises a particular relationship between a pair of words depends on two things: is he/she so familiar with the language in question that the concepts referred to can be properly identified, and does he/she have the sufficient knowledge about the concepts to recognise the relationship. For

instance, that *humerus* is an upper arm bone in the human skeleton is a fact. There is a relation of meronymy between *humerus* and *skeleton* in this respect. To recognise this relationship requires that the language user must be able to decode *humerus* to refer to the specific bone, i.e. the concept of HUMERUS and realise that *skeleton* though able to refer to the skeleton of any vertebrate, its most immediate reading is that of the HUMAN SKELETON. Furthermore it is necessary to have knowledge about anatomy to know that this particular bone is a part of the human skeleton.

This example, though a bit on the extreme, illustrates that apart from familiarity with the particular language of the language exchange, “world knowledge” possessed by the participants is important as well for the exchange to succeed. The sender makes assumptions on what he thinks his recipient knows about the world, and intuitively forms his utterance accordingly, that is if he cares whether the utterance is understood or not. Likewise the attending recipient intuitively tries to relate the utterance perceived to what knowledge he has about the world, and tries to make the content fit with this. The diagram below is adapted from memory and is to my knowledge a useful standard representation of circumstantial influences on linguistic form and style. It shows the four main ingredients in any linguistic interaction (the square boxes) and illustrates how these may influence each other. Together all of these notions have profound influence on the actual message sent by the sender, and how the receiver interprets it. Since I want to investigate mechanisms to realise the Topic of the “transmission”, I should try to establish a situation where all the other variables are locked into position in order to be able to concentrate on the Topic alone (since the way I use the term context in this project, relates most closely to Topic in the diagram).

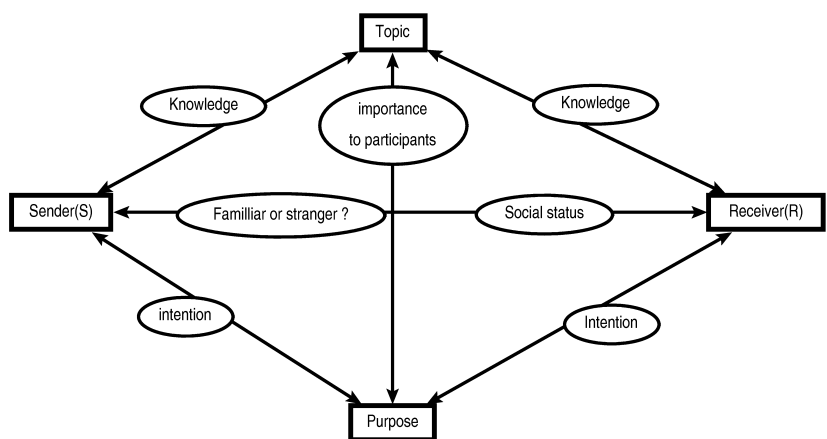


Figure 14: Detailed model of conversational influences. My experimental system will have one author, (i.e.: the Sender(S)), one purpose and one receiver, namely the skimming prototype. Only the topic should change over the course of the experimental corpus.

So for an exchange to succeed, active effort is necessary on both parts of the exchange. The lexical relationships described in the previous

paragraphs can be seen as tools for this effort. This being said, it is clear that they are not the only tools, but they are necessary none the less. For the lexical relationships to be used literally, it is also a requirement however, that there is an agreement of the purpose of the exchange. For informative tasks this purpose is to transmit information from the author to the reader. Under other circumstances it may be to entertain and surprise. There are also circumstances where the agreement is only a perceived one. It could be that the author deliberately wants to mislead the reader, this requires that the reader thinks he is being told the truth when in fact this may not be the case.

It is my intention to investigate the effect of lexical relations on the proper understanding of written text and this is to me best started by investigating the neutral case, where things are what they seem to be.

This is the reason I want to restrict my data to be seen as informative with the purpose of educating the reader. In this case it can be expected that the author has gone to some significant length to make his text as readily understood by the intended audience as possible. H.P. Grice supports this in his book ‘Conversational Implicature’ ,(Grice 1975), where he:

“... proposes a system of ‘conversational logic’ based on a number of ‘maxims of conversation’, i.e. intuitive principles which are supposed to guide conversational interaction in keeping with a general ‘co-operative principle’ (often referred to in the literature as CP). ‘Maxims’ differ from ‘rules’ in that they are seen as generally valid rather than to count only for specified (and specific) cases. The CP says:

- Make your contribution such as is required, at the stage at which it occurs, by the accepted purpose and direction of the talk exchange in which you are engaged (Grice 1975).” (Verschueren 1999).

Grice’s maxims are:

1. *The maxim of quantity*:
 - i) Make your contribution as informative as is required for the current purposes of the exchange.
 - ii) Do not make your contribution more informative than is required.
2. *The maxim of quality*: Try to make your contribution one that is true
 - i) Do not say what you believe to be false.
 - ii) Do not say that for which you lack adequate evidence.
3. *The maxim of relation* (later called *relevance*): Be relevant.
4. *The maxim of manner*: Be perspicuous.
 - i) Avoid obscurity of expression.
 - ii) Avoid ambiguity.
 - iii) Be brief.
 - iv) Be orderly. (Source: Verschueren 1999).

This principle has many consequences for how to formulate linguistic expressions to different purposes. Also it functions as a “rule-of-thumb” of how to interpret them. When speaking of the intentions of language interactors and purposes and circumstances for the language interaction, I have already plunged deep into the field of “pragmatics”. The problems addressed in this paper is to a large degree pragmatic ones, since they pertain to language use, that is how humans use language to achieve the intended goals. I will point out that I deem it essential for the design of any language semantics system, to at least “be aware” of what pragmatic mechanisms are at play in the data that the system is to analyse. Awareness like this will help guide design issues and decisions about interpreting conventions. Also, if a system is designed with the pragmatics in mind, it may very well help facilitating reuse of the system for different purposes within language processing. See for instance (Verschueren 1999), for an in depth treatment of pragmatics in natural language.

My system is for retrieving information from an intended informative text. The author intends to make this information as clear to the reader as possible. Furthermore the author expects the reader to be interested in learning this information. The reader, in this case my system, should therefore adopt the attitude that everything makes sense, so to speak. The game then, is to find out what sense it makes. This constitutes a situation where the cooperative principle of Grice is in its neutral position, where all maxims are intended to hold, and can be expected to hold.

In restricting myself to dealing with informative text only, I can therefore regard utterances as what they seem, in that I can expect the text to be as clearly formulated as possible by the author. In particular this means that when *humerus* and *skeleton* is mentioned in the same text or portion of text I can expect them to be related - despite the fact that one might take *skeleton* to refer to the carrying *backbone* of many different kinds of structures like bridges and houses for instance.

4 The machine-readable dictionary

The choice of machine-readable dictionary (MRD) imposes many vital decisions of design to the incorporating system. Most importantly the definition of the term “lexeme” discussed in the previous chapters is to a large degree dictated by the definition adopted by the MRD used. Therefore it is worthwhile devoting some effort in describing its design and how it works. The following excerpt is from the official web page:
<http://www.cogsci.princeton.edu/~wn/>.

“WordNet® is an online lexical reference system whose design is inspired by current psycholinguistic theories of human lexical memory. English nouns, verbs, adjectives and adverbs are organized into synonym sets, each representing one underlying lexical concept. Different relations link the synonym sets.

WordNet was developed by The Cognitive Science Laboratory at Princeton University under the direction of Professor George A. Miller (Principal Investigator).

Over the years, many people have contributed to the success of WordNet. At the present time, the following individuals at Princeton work on the continuing development of WordNet and applying it to research:

- O Professor George A. Miller
- O Dr. Christiane Fellbaum
- O Randee Teng
- O Susanne Wolff
- O Pamela Wakefield
- O Helen Langone
- O Benjamin Haskell

Dr. Fellbaum's work was supported in part by grant No. 9805732 from the National Science Foundation.”

The developers made WordNet’s core database freely available for everyone to download directly from their website for any purpose whatsoever, provided their copyright notice and disclaimer is provided as well. So here it is:

WordNet 1.6 Copyright 1997 by Princeton University. All rights reserved.

THIS SOFTWARE AND DATABASE IS PROVIDED "AS IS" AND PRINCETON UNIVERSITY MAKES NO REPRESENTATIONS OR WARRANTIES, EXPRESS OR IMPLIED. BY WAY OF EXAMPLE, BUT NOT LIMITATION, PRINCETON UNIVERSITY MAKES NO REPRESENTATIONS OR WARRANTIES OF MERCHANTABILITY OR FITNESS FOR ANY PARTICULAR PURPOSE OR THAT THE USE OF THE LICENSED SOFTWARE, DATABASE OR DOCUMENTATION WILL NOT INFRINGE ANY THIRD PARTY PATENTS, COPYRIGHTS, TRADEMARKS OR OTHER RIGHTS.

I will not describe WordNet in completeness but refer to the Aboce mentioned official homepage for further information. I will devote some effort to describe those parts of the resource that I deem necessary to understand the ideas of my system.

My experiment is to examine the relationships between the meanings of nouns and how these relationships may be exploited to disambiguate nouns and come closer to and understand of the contexts they occur in. The function of the MRD in this respect is to be the source of reference to the different meanings of each noun. What may be called ontological knowledge should also be readily available in the MRD.

4.1 Structure of the database

The WordNet database contains approximately 57,000 nouns organized into approximately 48,800 word meanings or synonym sets or senses. I used version 1.6 of the database, the current version is 2.0 and here the number of nouns have almost doubled. The hypernymous, meronymous and antonymous relationships between senses are also represented. The database format is as simple Prolog facts and as such WordNet is a suitable source for experiments of considerable size.

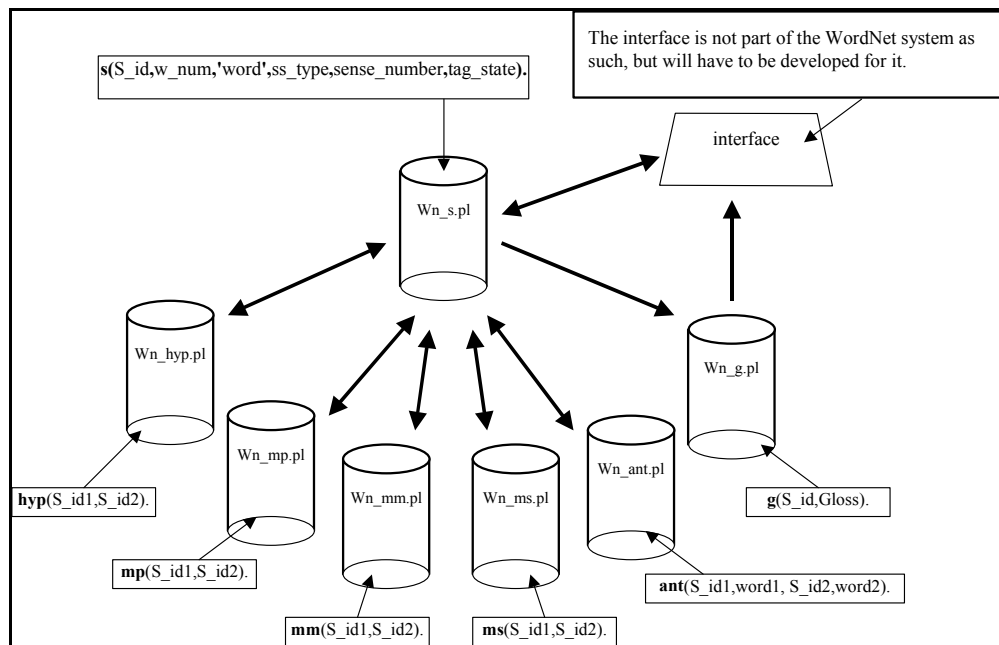


Figure 15: An schematic description of the structure of the WordNet database system. The figure shows the central synset-database, the hyponym-database, the three meronym-databases, the antonym-database and finally the glossary database. Associated with each of these is the format for their respective entries of facts. The transitions between databases represents how one may navigate through system.

WordNet consists of a series of separate databases of prolog facts. The pivoting item is the synset identification number (**Synset_id** or just **S_id**). To get the meanings of a word it will have to be sought out in the synset database, the synset_ids of the senses that can be realised by the word can then be assessed. To get a natural language explanation of those senses their synset_ids can be looked up in the glossary database. Similar operations can be made for the purpose of getting the immediate hypernyms of each sense or their meronyms. Note that the WordNet database does not come with an interface and such programming have to be done for the task at hand.

The way WordNet organizes its information is sought to facilitate rapid lookup times. The `synset_id`, though an abstract number, can be seen as a handle to the specific sense associated with words attributed with it. In this sense the `synset_id` represents what I have previously referred to using SMALL CAPITALS. `Synset_ids` are present in all corners of the WordNet database as it functions as the one unique key of the database as a whole.

4.1.1 An illustrated example

To illustrate how WordNet can be used an example seems in place. A look up in the synset database of the word *hammer* as input yields the following entries as output. For ease of reference I added the lower case indices, they are not part of the database.

```

a) s (100761424, 1, 'hammer', n, 9, 0) .
b) s (102749231, 2, 'hammer', n, 8, 0) .
c) s (102788624, 1, 'hammer', n, 2, 1) .
d) s (102788918, 1, 'hammer', n, 7, 0) .
e) s (102789071, 1, 'hammer', n, 6, 0) .
f) s (102789208, 1, 'hammer', n, 1, 1) .
g) s (102789360, 1, 'hammer', n, 5, 0) .
h) s (104132049, 2, 'hammer', n, 4, 0) .
i) s (105553636, 1, 'hammer', n, 3, 0) .
j) s (200970581, 1, 'hammer', v, 1, 1) .
k) s (201149879, 2, 'hammer', v, 2, 0) .

```

Figure 16: There is one entry in the synset-database for each meaning of a particular word. Here there are 9 noun senses - represented by the “n” as the fourth argument and 2 verb senses represented by the “v”.

So WordNet has eleven listed senses for the word *hammer*. Using the conventions already introduced, each of these entries correspond to what I would have represented graphically as:

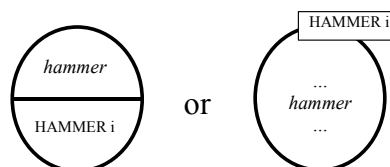


Figure 17: Two ways of depicting the same thing. The word *hammer* can have the meaning HAMMERi.

In the figures above the “HAMMERi” corresponds to the id numbers in each of the entries a)-k). So “100761424” would be for instance “HAMMER a)” and so on. The small “...” in the representation to the right indicates that indeed other words apart from *hammer* may realise this sense. Explicitly *hammer* may have synonyms or near synonyms (hence the `synset_id` seems a good name for the sense, since any sense can be seen as a set of (near-) synonymous word(s) realising it).

These entries are listed in the senses database, one of sixteen separate databases of WordNet. Going through the entries above each has the following format (from the original documentation):

`s(synset_id,w_num,'word',ss_type,sense_number,tag_state).`

Where:

- `synset_id` is a unique integer.
- `w_num` is the relative frequency of use with this word for this sense.
- `'word'` is the word in quotes.
- `ss_type` is the syntactical category of the sense (noun verb etc).
- `sense_number` is the relative number of this sense to this word.
- `tag_state` is a Boolean. `tag_state` is **1** if the sense number was assigned based on frequency of use, and **0** if it was not.

With this information we can see that all the senses of *hammer* above are indeed different, since the `synset_ids` are all different from each other. Near synonyms of *hammer* in each sense will share the `synset_id` but differ in `w_num`. As such near synonyms can be seen as members of the same set having the same `synset_id` or sense.

Furthermore we see that the last two entries refer to verbal senses because of the “v” in `ss_type`. It can also be seen that reference to the senses can take place in two ways, either with the `synset_id` as a whole, or with the pair of (`hammer`, `sense_number`).

We still don't get much information on what the different senses actually are, however. To this end WordNet supplies a glossary database that pairs each `synset_id` with an explanatory wording and sometimes also a clarifying example of use. The glossary entries for the different senses of *hammer* look like this, leaving out the verbal senses j) and k):

- | |
|---|
| <p>a) <code>g(100761424,'(the act of pounding (delivering repeated heavy blows); "the sudden hammer of fists caught him off guard"; "the pounding of feet on the hallway"))</code>.</p> <p>b) <code>g(102749231,'(a small mallet used by a presiding officer or a judge))</code>.</p> <p>c) <code>g(102788624,'(a hand tool with a heavy rigid head and a handle; used to deliver an impulsive force by striking))</code>.</p> <p>d) <code>g(102788918,'(a power tool for drilling rocks))</code>.</p> <p>e) <code>g(102789071,'(the felt-covered striker that causes the piano strings to vibrate))</code>.</p> <p>f) <code>g(102789208,'(the part of a gunlock that strikes the percussion cap when the trigger is pulled))</code>.</p> <p>g) <code>g(102789360,'(a heavy metal sphere attached to a flexible wire; used in the hammer throw))</code>.</p> <p>h) <code>g(104132049,'(the ossicle attached to the eardrum))</code>.</p> <p>i) <code>g(105553636,'(an athletic competition in which a heavy metal ball that is attached to a flexible wire is hurled as far as possible))</code>.</p> |
|---|

While this certainly helps a human reader, the natural language explanations are of very limited use to a computer. For a computer program to even come close to deciding autonomously which of the above sense to assign to *hammer*, the relationships of each alternative to other word senses occurring in the data has to be reasoned with. Both hypernymous and meronymous relationships are represented in WordNet as pairs of `synset_ids`, each specific type of relationship in its own separate database, and with its own operator as well. The immediate super-ordinates or hypernyms of the senses above are found as the second argument in the following facts from the hyponym database:

”**hyp**(`synset_id`,`synset_id`).

The **hyp** operator specifies that the second `synset` is a hypernym of the first `synset`. This relation holds for nouns and verbs. The reflexive operator, `hyponym`, implies that the first `synset` is a hyponym of the second `synset`.”

(from the official documentation)

```
a) hyp(100761424,100760227).
b) hyp(102749231,102961925).
c) hyp(102788624,102795523).
d) hyp(102788918,103168791).
e) hyp(102789071,103428699).
f) hyp(102789208,103428699).
g) hyp(102789360,103388599).
h) hyp(104132049,104096217).
i) hyp(105553636,105552741).
```

Glossary for these immediate hypernyms or super-ordinates:

```
a) g(100760227,'(a powerful stroke with the fist or a weapon;
"a blow on the head")').
b) g(102961925,'(a short-handled hammer with a wooden head
used to strike a chisel or wedge)').
c) g(102795523,'(a tool used with workers'' hands)').
d) g(103168791,'(a motor-driven tool)').
e) g(103428699,'(the part of a mechanism that strikes
something)').
f) Do.(same as above)
g) g(103388599,'(equipment needed to participate in a
particular sport)').
h) g(104096217,'(a small bone; especially one in the middle
ear)').
i) g(105552741,'(a competition that takes place on a field
rather than on a running track)').
```

Note that these new `synset_ids` are still senses that are realised by words, even though we have not explicitly made reference to any of these senses. To get possible realizations of these senses the Sense database

should be used again, this time using a `synset_id` as input and a list of words realising the particular sense as output.

It should be apparent that the distinctions made in WordNet is quite fine grained, and that differences in meanings are represented fairly faithfully. Following the graphical conventions that I have introduced in the earlier chapters the relationships would be represented like the following (note that the orientation of the arrows indicating hyponymous relationships have changed to vertical instead of horizontal. The convention is that the direction of this relationship is from subordinate to super-ordinate, and that subordinates are under their super-ordinate):

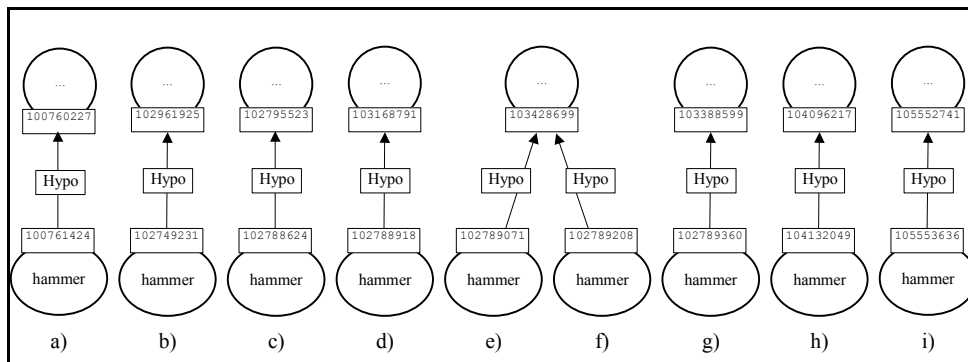


Figure 18: Graphic representation of the of looking up the meanings of the word *hammer* and finding the immediate hyponymous superordinates.

Referring to Figure 15 we can sum up that:

- Senses of words are found in the Synset-database (`Wn_s.pl`) along with the synonym relation.
- Synonymy is represented indirectly, so that synonyms have the same `Synset_id`.
- Lexical relationships (other than synonymy) are found in their own separate databases, in particular:
 - hyponymous relationships in `Wn_hyp.pl`.
 - antonymous relationships in `Wn_ant.pl`.
 - meronymous relationships in `Wn_mm.pl`, `Wn_mp.pl` and `Wn_ms.pl`.
- Glossary explanations of senses are found in `Wn_g.pl`.
- There are other relationships represented in WordNet, but they are particular to word classes other than the noun class, and so of little concern to the task at hand.

4.2 Some remarks to the use of WordNet

WordNet has several obvious benefits to a project like the one I am undertaking:

- WordNet is Free.
- WordNet has sufficient coverage to serve as a lexicon for real life data.

- WordNet is sufficiently fine grained to distinguish many subtle differences in meaning.
- WordNet attempts to describe the semantics of English on the grounds of lexical relationships as they are found to be in the minds of the native speaker of English.
- The format of WordNet makes it easy to experiment with.

All of these serve as arguments in favour of why I chose to use WordNet. However the following issue had to be addressed:

The member-substance meronymy relationship is represented in the database as:

```
ms(synset_id,synset_id).
```

However my prolog has a built-in two-argument predicate, `ms/2`, for measuring how many milliseconds a predicate takes to complete. To avoid the name clash the functor of this relationship was renamed in the database.

This was a very compact description of the format and functions of WordNet hopefully sufficient to understand the way I use this resource in my experimental prototype. The task with regard to WordNet is to write a suitable interface for accessing the information available in the database. The Words in the database are represented in their orthographic form, so this will have to be derived from the actual occurrences of the different forms prior to access.

5 The experimental corpus

2 "There are so many worlds, and I have not yet conquered even one."
INTRODUCTION
Five Impulses of Civilization
There is no single driving force behind the urge toward civilization, no one goal toward which every culture strives. There is, instead, a web of forces and objectives that impel and beckon, shaping cultures as they grow. In the Civilization III game, five basic impulses are of the greatest importance to the health and flexibility of your fledgling society.
Exploration
An early focus in the game is exploration. You begin the game knowing almost nothing about your surroundings. Most of the map is dark. Your units move into this darkness of unexplored territory and discover new terrain; mountains, rivers, grasslands, and forests are just some of the features they might find. The areas they explore might be occupied by minor tribes or another culture's units. In either case, a chance meeting might provoke a variety of encounters.

Figure 19: Raw corpus excerpt.

Formal linguistics is concerned with the design of formal representations for natural language structure and meaning, be it formal grammars, parsers or interpreting systems. Theories in this field should ultimately reflect actual language use as it occurs everyday between speakers of the language. For the obvious reason that it is impossible to formulate anything serious without access to massive amounts of data, the use of **corpora** is widespread throughout the field. A corpus in this respect is just a collection of actual linguistic data, either newspaper archives, transcribed conversation recordings or books. There are several advantages to using corpora, as there are some serious drawbacks.

The advantages include that considerable amounts of data can be achieved and analysed. To describe linguistic mechanisms formally and to verify the descriptions in any serious degree, a vast amount of data is needed to take heed of special cases and peculiarities that it is impossible to encompass with say a set of data constructed for the purpose. Another advantage is that some control is possible, with respect to the kind of language to analyse. Using a proper corpus, it is possible to focus analysis on restricted linguistic forms, like newspaper-articles, advertisements formal guides or everyday informal language to name but a few possibilities. As a remark, I want to point out that the experimental corpus used here is just that, i.e. experimental. Once a reliable prototype has been implemented and tested, it should be applied and tested on a much larger and broader corpus.

Among the disadvantages is a very serious risk of the data being outdated, since language changes rapidly (though often discretely), the time necessary to compile and publish the sources of languages may indeed secure that the data is outdated before it becomes available. The aim of the project is however, to investigate general semantic and pragmatic mechanisms and as such it does not matter if a few words have shifted their

meanings ever so slightly, as long as the underlying mechanisms remains the same.

A corpus does never the less represent real language, if nothing else, then at least at the time of the writing of the corpus. And as long as I do not tamper with the data it is also represents original natural language, and as such a suitable corpus can be a good source for research in actual language use. In order to make general points, however, I must try to pinpoint how the chosen corpus aligns itself with - or differs from language use in general.

5.1 Civilization III – advantages and problems

I have chosen the manual for the pc-game ©Civilization III (Infogames Interactive 2001), as the corpus for my project. I did so for several reasons:

- A manual represents informative text, where it is of obvious importance for the sender and intended recipient that the recipient understands the message of the text. This means that the language used in the manual can be expected to as clear and readily understandable as possible. In other words, either party can expect the other to do an active effort to make the linguistic exchange succeed, and in particular the cooperative principle can be expected to hold.
- The Civilization III game is a game of exploration, research, warfare, culture, economics and more. The developers tried to make it a detailed reflection of the entire history of development and expansion of human cultures on the grand scale. As such the manual can be expected to encompass a considerable variety of more or less related concepts and contexts, which is exactly what I want to explore.
- The manual is big enough to form a corpus of its own, and it is not necessary to incorporate other sources. This is an advantage in an experimenting project such as my own. (The finished system should of course be tried against differences corpora as well).
- Finally using a game manual rather than say a manual for a TV set has the possibility of adding a bit of colour to the task of reviewing output.

There are some problems with the chosen corpus that need to be addressed as well, however. :

- The high frequency of context/topic change may prove to be a problem since the guessing system might not be allowed enough text within one defined context to come up with a useable guess, before the context changes.
- The format of the original manual includes all kinds of "extra-linguistic" information like tables, illustration-comments,

flavour text ad so on, some decision have to be made in this regard.

- The language itself is running language in all its diversity. The manual is full of inflected words and sentence boundaries, as well as headlines. Some pre-processing is necessary before the word forms can be looked up in the MRD and related to each other.

Addressing the problems in order, the frequency of context/topic change is as much an advantage as it may pose a problem. There is not much sense in trying to recognise different context, if the corpus has no or very few changes in context. Using a corpus with a lot of different contexts, however, is bound to make problems and core issues apparent.

The tables and illustrations are not readily comparable to the linguistically mediated information available in the manual and thus will be ignored in the finished corpus. Illustration text and flavour phrases are however examples of language. More so they are examples of language use specific to the textual form. Taking into regard that the author must have included them for a purpose, and that this purpose quite likely is to better convey the intended illumination to the reader, these parts of the manual are included in the corpus where possible.

5.2 The tagger – TOSCA-ICLE

In order to be able to look up the nouns of a running text in the MRD, it is necessary to know what words are in fact nouns and what their dictionary form is. This morphological analysis I will have the external tagger TOSCA to handle for me.

The TOSCA system was developed at the department of Language and Speech at the university of Nijmegen (<http://lands.let.kun.nl>), and it is documented fully elsewhere (Haan and Halteren, 1997).

A tagging-system accepts running text as input and breaks it into constituent units, i.e. words and symbols. Furthermore it associates with each unit a tag, referring to the category of the unit as recognised by the tagging program. The TOSCA also analyses the units morphologically, and provides the orthographical forms (dictionary forms) of the words it encounter and recognise. TOSCA also recognises sentence boundaries.

There are a lot of taggers made available by universities and interest groups throughout the Internet, and most likely many of them would serve just as well. For my purpose I only use the facilities of the tagger that assigns an orthographical form and word class to each word form encountered in the data, and to a certain extent the markers for headlines as well.

5.2.1 Pre-processing

Time to look at some concrete data. The textbox at the start of this chapter contain a small excerpt of the civilization III manual. The excerpt is from the start of chapter 2 of the manual, and even without illustrations and tables, it should be easy to see that there are several kinds of text present.

```
2 "There are so many worlds, and I have not yet conquered
even one."
```

This part is a chapter number and a flavour text meant to prime the reader as to what to expect from the chapter.

```
INTRODUCTION
Five Impulses of Civilization
Exploration
```

These three lines are each examples of headlines. The all-capital headline is for head of chapter, where as the two others are examples of head of paragraph headlines. These will have to recognised and marked as such, since again the author may very well have chosen them for the purpose of clarifying his message. Indeed headlines may hold vital information for disambiguation of the relevant portion of text (chapter or paragraph). Information, that would be lost if the headlines was treated as just ordinary text in the paragraphs or chapters. Even though the present work will not make it quite to that point, I never the less want to separate headlines from body text, for the above reasons.

Furthermore in the data example, it can be seen that the style of the manual incorporates sentence boundaries with no intervening white space:

```
... strives.There ...
... exploration.You ...
... dark.Your ...
... find.The ...
```

These are notoriously difficult for computers to handle since the strings will often get interpreted as one unit instead of in fact three.

5.3 Solving the problems

I chose to use the UNIX stream editor, SED, to do my pre-processing. Some explaining is in place.

Headlines

Headlines in the civIII document comes in several flavours:

- Single words starting with an uppercase letter and no periods.
- Single words with all uppercase letters, possibly with a number and no period.
- Several words, all of which start with an uppercase letter and no period.
- Several words, all of which are all uppercase letters and no period.
- Several words, a few of which start with a lowercase letter and no period.

All of these are captured by the following regular expression patterns (the patterns are broken up for readability, see Unix documentation for further info on SED, for example (Loukides and Estabrook 1999).

This one takes care of the headlines consisting of several words. Starting with at least 1 uppercased word; possible followed by at most 3 lowercased words; this must be followed by at least one uppercased word, which is finally possibly followed by a number:

```
/
^
\[A-Z][A-z]*\)\{1,\}
\[a-z]*\)\{0,3\}
\[^\^a-z][A-z]*\)\{0,1\}\)\{1,\}
\[ [0-9]\{1,\}\)\{0,1\}
$
/
```

It also captures some undesired instances involving sentence boundaries these have to be filtered away using the following statement that no periods, commas etc. are in any headline:

```
/^[^.,;:]*$/
```

This one is much simpler taking care of headlines consisting of single words:

```
/^[A-Z][A-z]*$/
```

This suit of patterns captures almost each and every instance of a headline in the civIII document. There may be a few exceptions but not to an extent that will disturb the results in any significant way. Therefore they are, for now allowed to be there. I introduce a new line before and after each instance captured by the above patterns.

Sentence Boundaries

The separating space between sentences are introduced by the following SED-command, that replaces each instance of a period and an uppercase letter by a period, a space and that uppercase letter:

```
s/\.([A-Z])\.\ \1/g
```

These together constitute the following SED-script that will do my pre-processing.

```
# sentence boundaries fixed :
s/\.([A-Z])\.\ \1/g

# single word headlines fixed :
/^[A-Z][A-z]*$/ {
i\
a\
}

# multiple word headlines fixed :
/^[^.,:]*$/ {
/^\([A-Z][A-z]*\)\{1,\}\([a-z]*\)\{0,3\}\([\^a-z][A-z]*\)\{0,1\}\)\{1,\}\([0-9]\{1,\}\)\{0,1\}$/ {
i\
a\
}
}
```

After these transformations the data is sent through the pre-processing part of the TOSCA-system. Note how the new lines around headlines have caused the pre-processor of TOSCA to wrap them in <#>'s. Apart from separating headline from the rest of the text, this symbol helps the tagger to recognise that this is indeed a headline. Also note how Tosca counts "sentences" marking them by "**integer**".

The pre-processing is now done and the result can be passed to the tagging system itself. The tagger can be configured in many different ways to produce different layouts, I will however not document the tagging system in any detail, since it has been done elsewhere, and is as such not a part of my project. Part of the output of the tagger can be seen in Figure 21. In the beginning I will not distinguish headlines from the rest of the text, but I have made it possible to do so at a later stage in the development cycle of the system.


```

2 "There are so many worlds, and I have not yet conquered even one."
<#>
INTRODUCTION
<#>
<#>
                    Five Impulses of Civilization
<#>
There is no single driving force behind the urge toward civilization, no one
goal toward which every culture strives. There is, instead, a web of forces
and objectives that impel and beckon, shaping cultures as they grow. In the
Civilization III game, five basic impulses are of the greatest importance to
the health and flexibility of your fledgling society.
<#>
Exploration
<#>
An early focus in the game is exploration. You begin the game knowing almost
nothing about your surroundings. Most of the map is dark. Your units move
into this darkness of unexplored territory and discover new terrain;
mountains, rivers, grasslands, and forests are just some of the features
they might find. The areas they explore might be occupied by minor tribes or
another culture's units. In either case, a chance meeting might provoke a
variety of encounters.

```

Figure 20: After the initial pre-processing the excerpt looks like this.

In the first column of the output (Figure 21) is the actual word or symbol encountered in the original (pre-processed text). The second column holds the orthographical or uninflected form of the word, or a handle for symbols. Third Column holds the tag for the unit, complete with word class and inflective information of the interpretation chosen by Tosca as the most probable. The fourth column in this format holds the number of alternative interpretations recognised by Tosca. It is possible to adjust the system to give a list of alternative tags, which will indeed be necessary since The TOSCA tagger does make blatant mistakes in its judgements from time to other. In my work, the mistakes shouldn't influence on the general behaviour of the system due to its experimental character. Therefore the corpus will for now be the list of orthographic nouns as they occur in the Tosca output. It is easy to see how such a list can be derived from the Tosca output, so I will not burden myself with the task of actually implementing this simple rule governed algorithm, but make the experimental list by hand, simply copying each unit from the second column that has N(..) in the third column. I end each line with “.” (period) to make them syntactical Prolog facts.

^ 1			
2	2	NUM(card,sing)	
"	“	PUNC(oquo)	
There	there	EXTHERE	... (1)
are	be	VB(lex,intr,pres)	... (4)
so	so	ADV(connec)	... (4)
many	many	PRON(quant)	
worlds	world	N(plu)	
,	,	PUNC(comma)	
and	and	CONJUNC(coord)	... (1)
I	I	PRON(pers,sing)	... (3)
have	have	VB(aux,perf,pres)	... (11)
not	not	ADV(neg)	... (3)
yet	yet	ADV(ge,pos)	... (2)
conquered	conquer	VB(lex,montr,edp)	... (4)
even	even	ADV(ge,pos)	... (7)
one	one	PROFM(one,sing)	... (2)
.	.	PUNC(per)	
"	”	PUNC(cquo)	
^ 2			
<#>		MARKUP	
INTRODUCTION	introduction	N(sing)	
^ 3			
<#>		MARKUP	
<#>		MARKUP	
Five	five	NUM(card,sing)	
Impulses	impulse	N(plu)	... (5)
of	of	PREP(ge)	... (1)
Civilization	Civilization	N(sing)	... (3)
^ 4			
<#>		MARKUP	
There	there	EXTHERE	... (1)
is	be	VB(lex,intr,pres)	... (3)
no	no	PRON(neg)	... (3)
single	single	ADJ(ge,pos)	... (5)
driving	drive	VB(lex,intr,ingp)	... (5)
force	force	N(sing)	... (13)
behind	behind	PREP(ge)	... (4)
the	the	ART(def)	... (1)
urge	urge	N(sing)	... (12)
...

Figure 21: Sample output from the tagger. (The file was truncated for presentability).

The entire portion of the civilization manual that I used for experimental corpus, and the corresponding noun lists can be found in the appendix. The noun list of the excerpt used as example in this chapter is here. This is what is to be investigated for internal lexical relationships, to see if the topic(s) or domain(s) of the original text can be refined or at least hinted at automatically.

impulse.	web.	importance.	exploration.	terrain.	tribe.
civilization.	force.	health.	game.	mountain.	culture.
force.	objective.	flexibility.	surroundings.	river.	unit.
urge.	culture.	fledgling.	map.	grassland.	case.
civilization.	civilization.	society	unit.	forest.	meeting.
goal.	game.	exploration.	darkness.	feature.	variety.
culture.	impulse.	game.	territory.	area.	encounter.

Figure 22: It is easy to see how this list of nouns can be extracted from the tagger output.

6 Putting it all together.

Having described all the theoretical background it is now time to make explicit what my contribution is going to be.

Picture a person with a book. The person has not yet read the book but is going to do so. Usually, chances would be that already at this point he or she would have an idea of what the book is going to contain since the book has a title and people usually have a purpose in reading a book. Perhaps it holds answers to questions that the reader wants answered - instructions of how to solve a particular problem that the reader wants solved – or perhaps the purpose is of a more entertaining nature.

In the pictured case, however, the reader is being told that the book is informative but has, beyond that, no idea whatsoever of what can be expected from the content of the book. It would be as if the person was given the book to read and absorb just because someone else wants to be able to question the person about its content. Furthermore the person has never read a book before or experienced anything at all, he is a completely blank slate with regard to the world around him. To help the understanding of the book, the person has only a dictionary at his disposal, containing word meanings and relations between concepts.

This is a rough description of the tasks and conditions of the system I am trying to approach.

In fact my system is further restricted to read only the nouns of the book and to disregard everything else. Nouns, that may or may not be essential to the content of the book – nouns, that may indeed be ambiguous and misleading as well.

It should be clear that these very restrictive conditions do not allow for a fully detailed understanding of the content of the book. The “understanding” can at most be a vague summation of concepts that seems important to the content of the book and how these concepts may be related to each other in the small portion of “reality” described by the book.

If the reader were indeed a normal person, then the task would be trivial and not very interesting. - Perhaps mildly entertaining as a party game:

“Here is a list of nouns as they occur in one particular text. Regard them one for one and see how precisely you can gradually infer what the text is about! The first word is”

The skills needed for solving the problem are undoubtedly so fundamental to human perception - and indeed so closely related to the way we experience the world around us - that they look like innate abilities. I think it is fair to say that people are generally not even consciously aware of applying them - or how they apply them?

Here the reader is a computer program, however, and the task is not trivial at all. To me, the needed skills are so critically fundamental to language understanding that I think they must be explicitly modelled or

mimicked if we are ever to equip computer programs with even the most basic of semantic abilities. Without this we will not succeed.

No skills are obtained instantly, though, but rather achieved through training and experience in using the proper tools of the trade. To model these skills we must first try to understand what is the nature of the associated tools. Then those tools will have to be modelled as closely as possible before they can be experimented with. At this point it can be decided if more tools are necessary or if the ones that were modelled need refining. Finally, a training system can begin training with the tools (This paper will not deal with training at all. The tools modelled in this paper are however undoubtedly important in a trainable system).

To understand what tools are necessary, perhaps it is useful to return to - and elaborate on - the “party game” of finding the links between the nouns of a text, that reveals what the original text is about.

6.1 The word game – an illustrated example

For the purpose of this descriptive example, I will digress from the CIVIII corpus for a bit and use an excerpt from a novel written by Stephen King, (King, Stephen 1987), as data. This I do, because the basic relationships between concepts are more apparent when using a fictitious text. Because the author describing a non-existing universe must describe it in more detail than if he could depend on the reader to draw on common knowledge to make deductions on the finer details. He is describing the world a new (well, he is describing a new world to someone who doesn't know anything about it yet).²

Indeed it will not take a person very long to solve the puzzle. A few minutes or maybe a little more should be enough to come up with a

“Regard these nouns one for one and see how precisely You can gradually restrict what the text is about, These nouns were taken from the first chapter of the novel by Stephen King, “The Eyes of the Dragon”					
Kingdom	roland	year	year	son	king
delain	king	face	god	peter	magician
king	land	hand	grace	man	
son	evil	court	kingdom	roland	
delain	work	year	baron	son	
kingdom	king	king	serf	thomas	
king	death	plaza	wife	king	
time	time	foot	king	man	
historian	heart	needle	roland	flagg	

Figure 23: A word game. The word list resembles closely what the computer program will take as input. The computer program will however accept the words one at the time and not have them all available from the start.

quite detailed picture of what the text must be about, that the nouns in Figure 23 were taken from.

The network in Figure 24 is an example of how such a picture could possibly look like. It was made by hand, arranging each noun with respect to what made the most sense to me. I included an edge between two nouns where their most likely interpretations seemed to be somehow related. This

² Also I will leave the Stephen King example as soon as I have shown my points.

is of course a very intuitive approach but it never the less depicts a detailed sketch of and inspiration as to what a coherent representation of the context of the text could look like, even though the system's guess need not be nearly as detailed as this, to demonstrate its use.

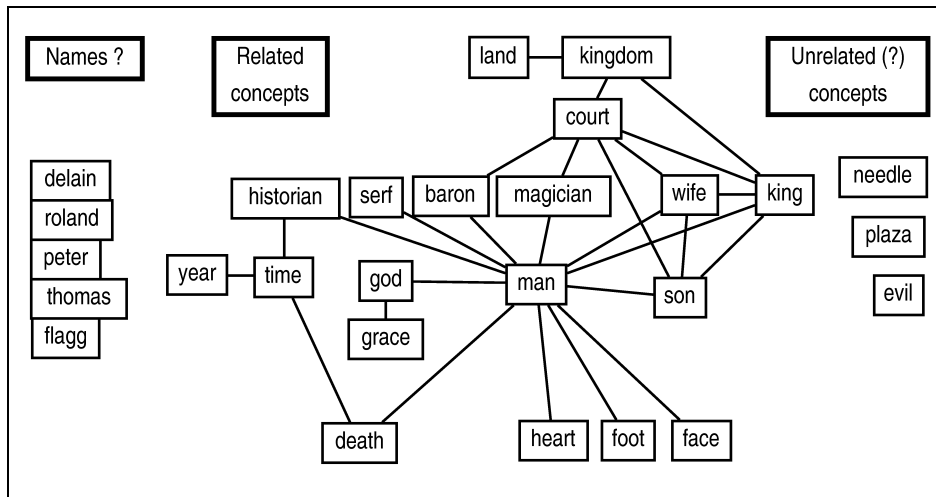


Figure 24: “The text that the nouns were taken from must be about a certain kingdom – a certain land. Its royal family and the court it keeps. Also there is something about mankind in general, and of life, death and religion. The names are probably referring to the central characters of the fairytale. There could be some evil opposing the kingdom? “ This could be the reasonable guess of a person presented with the nouns of Figure 23. The intuitive relationships between concepts - represented by edges in the network - are intended to mean: “A is somehow related to B” and vice versa.³

The network in Figure 24 illustrates two important points about the notion of context that I have been working with.

Firstly note how the edges in the network pose limitations on the reference of their ending points. It is most obvious when concerning the vertex containing the noun *man*. This noun is obviously polysemous, meaning both an adult male human, and also refers to mankind in general regardless of gender and age. The edges connecting *man* to other concepts in the context like wife, son, heart, foot, face make it clear what concept is referred to by the noun in this context, namely that *man* refers to - or rather realizes - the concept MANKIND.

Secondly, representing context as a network illustrates not only that the concepts are connected but also that they are indeed representatives of a small subset of all concepts in existence. A context is that small part of “everything” that is relevant to the message being conveyed or to the situations being described. The context can indeed be seen as closely related to the concept of “common knowledge”, that is, knowledge about the world necessary to correctly interpret the linguistic message in question. When designing “common knowledge” or “world knowledge” –bases for use

³ The relationships in Figure 23 include some that need explanation. The vertices “wife” and “son” are taken to mean the family of the “king”. Thus being the queen and prince they are both part of the “court”. The rest should be fairly self-explanatory.

within artificial intelligence it has been suggested to represent this body of knowledge using semantic networks (Nilsson, Nils J. 1998).

Indeed I do not think that is too far fetched to postulate that we, as human beings, identify concepts around us in terms of how they relate to the rest of the concepts in the world. It is very hard indeed to explain the meaning of a noun without reference to the main category of the intended concept, the purpose of its use, how, of what materials or by what is was made. (For very similar ideas and excellent discussion, see (Pustejovsky, James 1995)).

This is the main reason I choose to represent the possible solutions to the word game as a kind of semantic networks, having concepts in the vertices and have the edges express relationships between concepts.

Another beneficial consequence is that it makes the context in question useable as a mini dictionary; making it possible to distinguish specific knowledge from universal knowledge. Specific knowledge pertains perhaps only to the particular situation being described by the text, but we need not forget about the rest of the universe in the process. By this I mean that it could make sense to look in the context first for a meaning of a noun, and only when unsuccessful or in doubt, then turn to the dictionary for enlightenment, expanding the context as necessary to encompass the situation described by the text.

But how can this game be realized in a computer program? Basically the system should (like a human reader) treat the noun list sequentially from start to end one noun at a time. The “treatment” must fall in two phases, recording and interpreting. The recording requires a structure for storing facts related to the text/noun list being analysed. The interpreting requires a methodology for reasoning with the stored information.⁴

Once finished, the system should have a representation of that interpretation of the list (i.e. those interpretations of the nouns in the text that “make the most sense” or “are the most consistent”) and the way they might relate to each other. Perhaps it should also have a way of representing tricky interpretations – interpretations that “are less certain” so to speak. This representation could very well be in the form of a semantic network. Since there may still be more than one likely interpretation, one semantic network should be available for each “likely” interpretation. Summing up it looks like we must model following tools or abilities :

- 1) Ability for to read and hold the textual data for future reference.

⁴ The system should ideally alternate between these two phases in order to achieve a stepwise-refined “understanding” of what has been read so far, and using this “understanding” to further help the understanding of what is to come. The prototype being developed in this paper will record first a fixed scope (one sentence, one paragraph or one chapter), and only then interpret the recordings. Interpreting one paragraph with respect to the interpretation of the previous paragraph will be discussed in a later chapter but will not be implemented.

- 2) Ability to generate interpretations of the data as stored – and to distinguish different interpretations from each other.
- 3) Ability to measure and compare the “quality” of different interpretations.

6.2 Problem context

For the purpose of the following description of the problem and its attempted solution it will be useful to have the context of the problem formally stated.

6.2.1 Sets of the dictionary :

At the core is the MRD – in my case the WordNet database collection. Recall that the synset database of WordNet involves sextuples:

$s(m,Int,w,POS,n,Boolean)$, where

“m” is the synset_ID (this we can now treat as an abstract representation of the concept being realized by the words in the set of synonyms – their common “meaning”).

“w” is a word that can realize “m”.

“Int” reflects the “likelihood” of “w” actually meaning “m” and the Boolean at the end records whether this was decided based on frequency or arbitrarily. I do not use this measure in my system but refer to chapter 4 for a more detailed description of the MRD.

“POS” ranges over representatives of the different Part-Of-Speech categories, i.e., n for nouns, v for verbs etc.

“n” acts as a distinguishing mark of each word within the set of synonyms.

Instead of using the entire sextuple just described I will for clarity use a triple when referring to an entry in the synset database :

$s(m,w,n)$.

NOTE: As a confusing consequence of WordNet’s use of “n”, any one lexeme can be referred to in a total of three ways, namely (m_i,w_j,n_k) , (m_i,w_j) or (m_i,n_k) . Furthermore, please note that the triple, $s(m,w,n)$, properly identifies and represents the corresponding sextuple. I.e. we will not see a verb realizing a nominal concept or vice versa.

Other important parts of the WordNet databases include the three separate databases for hyponymy, meronymy and antonymy. The entries in

these databases respectively conform to patterns $h(m_i, m_j)$, $m(m_i, m_j)$ and $ant(m_i, n_k, m_j, n_l)$.

NOTE: Antonymous lexemes are represented in WordNet using the corresponding n's instead of the more natural w's

Finally, I will adopt the Prologian anonymous variable, “_” (underscore), to represent any variable where the value of the variable does not matter. Also remember that all words w in the MRD are the orthographical representative of their respective paradigms, i.e., there are no inflected word forms in the MRD.

Now the following sets can be distinguished :

W : The set of all words in the WordNet dictionary.

M : The set of all meanings or concepts in the WordNet Dictionary.

R : The set of all relations in the WordNet Dictionary – or rather : all instances of hyponymy, meronymy and antonymy in WordNet.

The words in W are for our purposes restricted to nouns only and each member of the set will be referred to as w - with or with or without an index, w_1, w_2 etc. Since W is a set, indices to members of W (or any subset of W) are intended as individualising marks and do not indicate any internal order. Members of W are the objects I have been referring to using *italics*. They would occur in the upper half of a Saussurian sign.

- $w \in W$ iff $s(_, w, _)$ is an entry in the synset database of WordNet (and w is a noun !).

The meanings in M are likewise restricted to be nominal concepts only, i.e. those concepts referred to by nouns. Each member of M will be referred to by m, m_1, m_2 etc. Again, since M is a set, indices to members of M (or any subset of M) are intended as individualising marks and do not indicate any internal order. Similarly, there is no logical connection between the indices of words and the indices of meanings. Meanings are the objects I have been referring to using SMALL_CAPITALS. They would occur in the lower half of a Saussurian sign.

- $m \in M$ iff $s(m, _, _)$ is an entry in the synset database of WordNet (and m represents is a nominal concept !)

In terms of M and W any pair of words and meanings, (w, m) , - such that $s(m, w, _)$ is an entry in the synset database - is what I refer to by “a lexeme”.

The relations in R comprise our selected set of binary relationships between nominal concepts of M . These relationships will be referred to as $Rel(m_i, m_j)$, where Rel is a variable ranging over relation names h for hyponymy and m for meronymy.

Also in R are the relationships of antonymy between nominal “lexemes”. The antonymous relationships will be referred to by

$a((w_i, m_i), (w_j, m_j))$). With the data structures of WordNet (again, refer to chapter 4 for details on the different data structures of the MRD) this is equivalent to $ant(m_i, n_i, m_j, n_j)$ as it appears in WordNet's antonymy database where $s(m_i, w_i, n_i)$ and $s(m_j, w_j, n_j)$ are the corresponding entries in the synset database. We get :

- $h(m_i, m_j) \in R$ iff $hyp(m_i, m_j)$ is an entry in the hyponym database of the WordNet Dictionary.
- $m(m_i, m_j) \in R$ iff
 - $ms(m_i, m_j)$ is an entry in the substance meronym database of the WordNet Dictionary OR
 - $mm(m_i, m_j)$ is an entry in the group meronym database of the WordNet Dictionary OR
 - $mp(m_i, m_j)$ is an entry in the part meronym database of the WordNet Dictionary.
- $a((w_i, m_i), (w_j, m_j)) \in R$ iff $ant(m_i, n_i, m_j, n_j)$ is an entry in the antonym database of the WordNet Dictionary AND $s(m_i, w_i, n_i)$ and $s(m_j, w_j, n_j)$ are both entries in the synset database.

6.3 The task

Within this setting the system should accept a sequence of nouns and provide a set of "interpreting graphs". Each graph will represent one "good partial interpretation" of the nouns contained in the text. Each graph must satisfy the following condition :

- there are no ambiguous readings, i.e. no noun has more than one meaning within any given interpretation.

The task is :

GIVEN: A sequence S of nouns " s_1, \dots, s_n ".

TO FIND: A set of "good" partial interpretations of S.

6.3.1 Implications

Input to the system will be a sequence of nouns in their dictionary form ordered as they occur in the original text. To be more to the point: There may be examples of nouns that occur more than once in the sequence. With respect to S, the indices represent the relative position of individual instances in the sequence. It will become handy to be able to distinguish between instances, words and meanings :

DEF: $S = "s_1, \dots, s_n"$, i.e., the input sequence of word instances.

DEF: $W_0 = \{w \mid w \text{ occur in } S\}$, i.e., the words.

DEF: $M_0 = \{m \mid s(m, w, _) \text{ is in the synset database AND } w \in W_0\}$, i.e., corresponding possible meanings.

The above definitions formally state that I will refer to the set of nouns that occur in S, i.e., without repetitions, as W_0 , i.e., $W_0 \subseteq W$. Similarly I will refer to the set of meanings that can be realized by nouns in W_0 as M_0 , i.e., $M_0 \subseteq M$.

Technically a complete interpretation amounts to defining a function, $I:W_0 \rightarrow M_0$. I will present a theoretic algorithm for finding such complete interpretations. A complete interpretation, however, requires that each and every noun in W_0 be disambiguated and I will show that this is not feasible for practical purposes, nor is it necessary for getting a general picture of what the text is about.

I will introduce the notion of a partial interpretation, $I^p:W_c \rightarrow M_0$, where $W_c \subseteq W_0$, concerning those nouns that the system did assign meanings to (note, that the associated set of meanings would be called M_c , i.e., the concepts of that particular interpretation, something that should correspond very closely to the semantic domain of the original text, if the interpretation is indeed “good”). The partial interpretation will then be measured by some metric of “goodness”.

I will illustrate the large-scale structure of the search by showing a non-deterministic first version of the algorithm. It begins with an empty partial interpretation, and successively assigns meanings to some of the words in the text, the order being directed by the degree to which word meanings are interrelated (according to the MRD).

Later, I will show how the non-deterministic algorithm can be refined to make it deterministic, and how to deal with multiple interpretations via a scoring scheme.⁵

The output of the system will be in essence threefold:

1. The first part will be a graph illustrating the lexemes of the interpretation (consisting of the nouns of W_c and corresponding meanings from M_c) as vertices, and the involved relations (hyponymy, ... etc.) along the edges.
2. The second part of the output from the system will comprise the list of those words from W_0 that were not assigned a meaning by the interpretation, i.e., $W_0 \setminus W_c$. Each of these nouns will be paired with their entire original set of alternative meanings because they are still ambiguous, since the system did not choose between their alternative meanings. To keep the often very long list representing this second part from cluttering up the picture, I will often avoid showing it explicitly. This being said, it is important to be aware of the second part as an important part of the systems answer to the puzzle.
3. As the final part of the output will be various statistics of the interpretation including the rating of the solution as computed by the scoring scheme.

⁵ In fact, things went in a slightly different order. The rationale on scoring was done while examining the complete interpretations, and the development of the algorithm for partial interpretations was done on the grounds of the result of that rationale. I am however not documenting the thought process, but rather its outcome.

6.3.2 On Goodness

To sum up and clarify, the ideal solution would be to find that one interpretation of the text that the author originally intended. That is however not feasible since even human readers “go wrong“ now and then, hence a computer program cannot be expected to do better. Also - most texts just are inherently ambiguous, even strictly informative ones. (Again, a creative interpretation of Grice’s Cooperative Principle could indicate that either A: Any of the alternative interpretations adequately represents the intended meaning, or B: The ambiguous parts of the text are of no - or only limited importance to the intended meaning of the text.) With this in mind, we will have to settle for a “good” interpretation. Finding a solid measure of “goodness” in this respect is the real task. I will deal with this task when discussing scoring later in this chapter.

6.4 My solution

As stated in the previous section, I will start out with a non-deterministic algorithm that for each input sequence produces exactly one solution. I will demonstrate that each possible solution produced by the non-deterministic algorithm satisfies my conditions for goodness and unambiguity (i.e., $s_i=s_j \Rightarrow I^p(s_i)=I^p(s_j)$). In later sections, I will show how to extend it to handle multiple solutions and rank them with respect to “goodness”.

The solution that I devised for the task falls in two steps :

1. Representing the data. In this step the system reads the input data sequence and makes all the necessary MRD-lookups. The resulting information is organised in the structure that I call the “CHUNK”. This allows for all further work to be done on the CHUNK and secures that time consuming MRD-lookups are done once for all.
2. Interpreting the CHUNK. In this step the system reasons with the information in the CHUNK and provides a partial interpretation to the original input noun sequence. The second step also computes the score of the partial interpretation and various statistics.

6.4.1 The CHUNK.

The CHUNK is in essence a structure for storing that portion of the MRD that applies to the current input data sequence. It will hold two main structures, RVOCAB and LINKS :

1. RVOCAB (running vocabulary of the system). The set W_0 of nouns, each member paired with the corresponding set of alternative meanings for that noun. Formally each of these pairs can be represented as $(w_i, M_0(w_i))$, $M_0(w)$ is defined below.
2. LINKS. The set, $R_0 \subseteq R$, of relationships between
 - a) Meanings, m , if the associated relationship variable is “hyp” or “mero”, and

- b) Lexemes, (w,m) , if the associated relationship variable is “ant” (later “syn” as well). See the definition of R_0 below.

The CHUNK as the above-defined pair of sets - and this is the way it will actually be represented internally by the system - can also be defined as a graph. Doing so perhaps demonstrates the relation between the problem and its solution better than the pair of sets. For this purpose let us define :

DEF.: $W_0(m)$ is that set of nouns from W_0 that can realize the meaning m . Its transposition, $M_0(w)$ is that set of meanings from M that can be realized by the noun w .

DEF.: CHUNK Graph. A CHUNK C is visualised as a labelled undirected graph $CG(V,E)$ with :

- a) A set V of vertices one for each $m_i \in M_0$. A m_i vertex has as label the set of nouns $W_0(m_i)$. (Note that I want to represent meaning-based partial synsets as the vertices of the visualisation, therefore the noun-based RVOcab is in essence transposed).
- b) A set E of labelled undirected edges - one for each member of LINKS, with the associated relational constructor (“hyp”, “mero” or “ant”) as the label of the edge. (Important Note : The current state of the graphic representation of the CHUNK distinguishes links between lexemes from links between meanings solely via the label r_k of the corresponding edges. Since the vertices may hold several different words an antonymous edge between a pair of vertices does not make explicit which pair of lexemes are in the antonymous relation to each other).

With respect to LINKS and R_0 , let R_M be the set of relationships between *meanings* in R - hyponymous, meronymous Furthermore let R_L be the set of relationships between *lexemes* in R – antonymous relationships in particular. R_0 can now be expressed by the following expressions:

$$R_0 = R_{M0} \cup R_{L0}, \text{ where :}$$

$$R_{M0} = \{ \text{Rel}(m_i, m_j) \in R_M \mid m_i \neq m_j \wedge \exists w_i, w_j \in W_0 \mid w_i \neq w_j \wedge m_i \in M(w_i) \wedge m_j \in M(w_j) \}$$

$$R_{L0} = \{ \text{Rel}((w_i, m_i), (w_j, m_j)) \in R_L \mid m_i \neq m_j \wedge \exists w_i, w_j \in W_0 \mid w_i \neq w_j \wedge m_i \in M(w_i) \wedge m_j \in M(w_j) \}$$

6.4.2 An Example:

Consider the following example to illustrate the entire design of the algorithm:

$$S_1 = \text{”}w_1 w_2 w_3\text{”} .$$

Note, that since all of the instances in S are different words we have a one-to-one mapping from S_1 to W_0 :

$$W_0 = \{w_1, w_2, w_3\}$$

Assume

- that each word has two alternative meanings,
- that two of the words, w_1 and w_2 , share a meaning, m_1 , between them.

Resulting in five distinct meanings,

$$M_0 = \{m_1, m_2, m_3, m_4, m_5\},$$

and six distinct lexemes,

$$\{(w_1, m_1), (w_1, m_2), (w_2, m_1), (w_2, m_3), (w_3, m_4), (w_3, m_5)\}.$$

The above set of lexemes can also be represented in terms of $M_0(w)$, (i.e., essentially the transposition of the corresponding synsets of WordNet)
:

$$M_0(w_1) = \{m_1, m_2\}$$

$$M_0(w_2) = \{m_1, m_3\}$$

$$M_0(w_3) = \{m_4, m_5\}$$

The set can also be expressed in terms of $W_0(m)$, (i.e., similar to the synsets of WordNet):

$$W_0(m_1) = \{w_1, w_2\}$$

$$W_0(m_2) = \{w_1\}$$

$$W_0(m_3) = \{w_2\}$$

$$W_0(m_4) = \{w_3\}$$

$$W_0(m_5) = \{w_3\}$$

Furthermore assume that there are hyponymous relations between meanings: $R_0 = \{\text{hyp}(m_1, m_4), \text{hyp}(m_3, m_5)\}$. The corresponding CHUNK, C_1 , is :

$C_1 = (\text{RVOCAB}, \text{LINKS})$, where

$$\text{RVOCAB} = \{(w_1, \{m_1, m_2\}), (w_2, \{m_1, m_3\}), (w_3, \{m_4, m_5\})\},$$

$$\text{LINKS} = \{\text{hyp}(m_1, m_4), \text{hyp}(m_3, m_5)\}.$$

The CHUNK serves as an abstract representation of the problem instance at hand and, as it often is, within a problem reside its solution(s). I will now show how this pertains to the CHUNK and its interpretation(s).

6.4.3 Interpreting the CHUNK

By now it should be clear that a complete disambiguation of the CHUNK implies finding exactly one lexeme for each of the words in W_0 , i.e., reducing the six lexemes from the above example to three. This can obviously be done through several different interpretations, each of which acts as a kind of filter to the CHUNK.

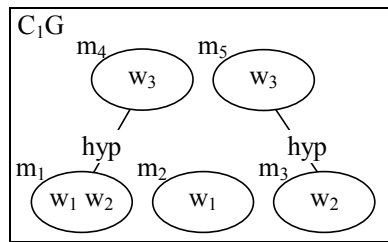


Figure 25: C_1G , the visualization of the CHUNK, C_1 , for the sequence, S_1 . Note how the vertices are named according to M_0 . Each vertex represents a set of nouns from W_0 that can realize the associated meaning, i.e., $W_0(m_i)$. Edges connect some of the vertices according to the lexical relationships between the associated meanings. The label of each edge indicates the nature of that particular relationship. As an example of antonymy, assume that the label of the left edge reads “ant” instead of “hyp”. Then the lexeme (w_3, m_4) would be the antonym of exactly one of the lexemes (w_1, m_1) or (w_2, m_1) . Which one of these however, cannot be shown in the current visualization of the CHUNK.⁶

Intuitively a complete interpretation including relevant lexical relationships can be established informally in three easy steps:

- 1) First, make a copy of the CHUNK where all edges and all members of all sets, i.e., the label of each vertex, are removed.
- 2) Then copy each word in exactly one of its original positions.
- 3) When each word appears in the label of exactly one vertex, copy those edges from the original CHUNK that have both ending points in vertices with nonempty labels.

Each word, w_i , in W_0 multiplies the number of possible interpretations by a factor equal to the number of members in $M(w_i)$. With regard to the sequence S_1 - introduced in 6.4.2, we get, $2*2*2$, eight different interpretations of C_1 as it is, since each of the three words can realize two different meanings. Figure 26 shows the eight different interpretations of C_1 . In each interpretation the irrelevant concepts, i.e.: those meanings that have no words realizing them, are shown in a shaded hue. Likewise any edge in the C_1 that would connect a shaded vertex in one particular interpretation is omitted in that interpretation

It comes as no surprise – and this example shows it rather clearly – that the number of interpretations to a given sequence is exponential to the length of the sequence.

Concretely, the worst-case will occur when all words in the sequence are different, i.e., S and W_0 are of equal length, namely n . I will assume the existence, at any given point in time, of a constant K representing the maximal number of meanings associated with any noun in the MRD. The worst-case scenario implies that all n nouns have that many meanings, i.e.: M_0 can contain up to $n*K$ different meanings. It follows that the number of different complete interpretations of any given string is bounded upwards by

⁶ The graphical chunk is however just a visualization and it can, no doubt, be revised to illustrate this kind of phenomena as well (for example underscoring the antonymous lexemes). Internally in the system there is no doubt about which lexemes are in the antonymy relation to each other.

K^n . This is quite an intimidating measure indicating that an algorithm intended to compare all the interpretations and find the optimal and complete among them, will inevitably have its complexity in the magnitude of $O(K^n)$. In order to circumvent this obviously intractable complexity it is necessary to realize that indeed we are not interested in all the interpretations but merely the “good” ones. The task then becomes to find a method for arriving at the “good” solutions without having to compare the lot of them.

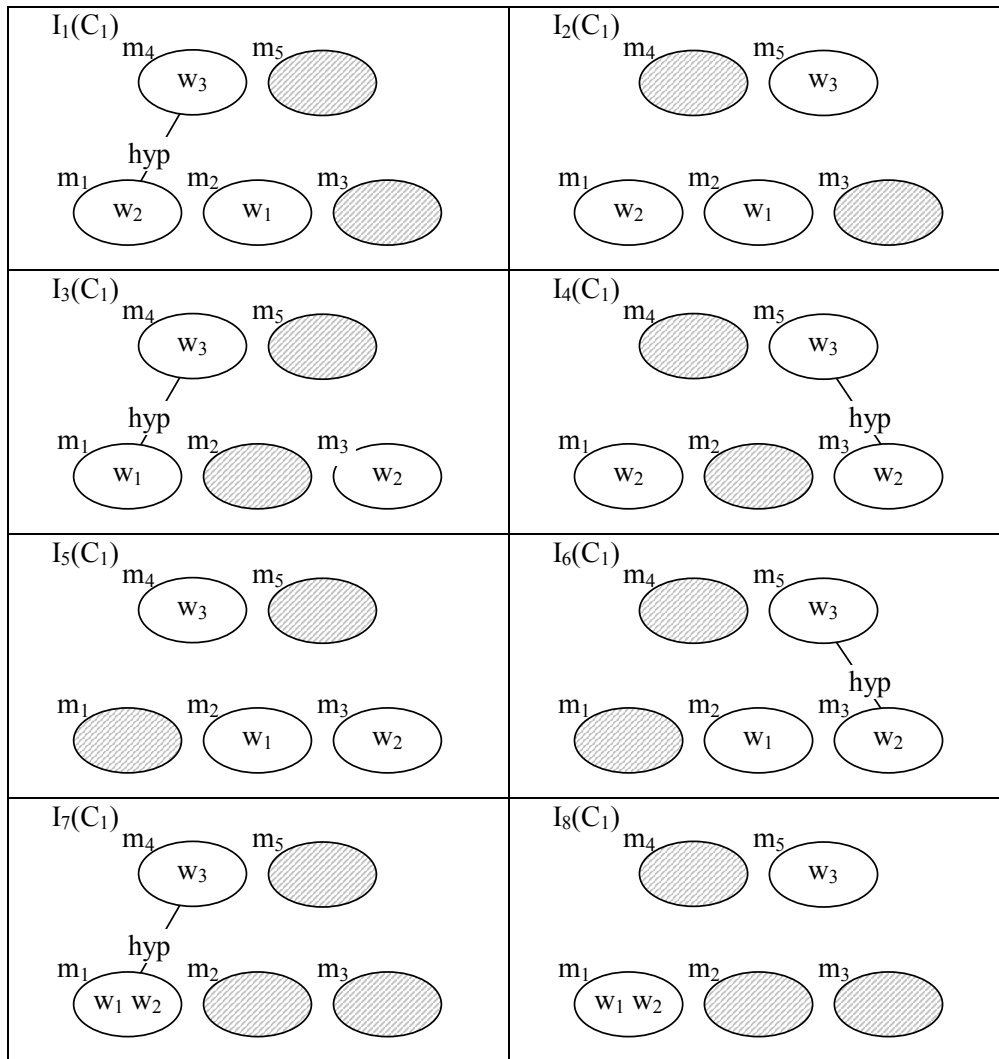


Figure 26: Each interpretation commit to certain lexemes of the CHUNK while it rules out others. The total number of different interpretations of a sequence is exponential to the number of instances - unless no word in the sequence is polysemous...

6.5 Partial interpretations:

This notion is spurred from the fact that even to humans some words just are ambiguous, no matter how much effort is invested in their disambiguation. People are not put off so easily, instead we insist on making as much sense out of the mess as possible, and relying on our world

knowledge and common sense to deduce the rest. Consequently it makes sense to disambiguate what we can and “hope” that the rest is either not “that important” or that any of the alternative interpretations of them really are equally sufficient for a useful general understanding of the text.

Under the assumption that lexical relationships between lexemes in the text are of substantial importance in the recognition of the proper semantic context of the text, my plan is to focus on the edges of the CHUNK instead of on the vertices. Because natural ontologies are generally very sparse structures, i.e., the number of edges is low with respect to the number of vertices, I expect an improvement in the performance of the resulting algorithm over the intractable one for complete interpretations.

6.5.1 The algorithm, informally and explicitly

My choice is that edges - and their respective ending points, from the CHUNK will be included in the partial interpretation one at the time. Each time an edge is included, the words that can realise either ending point will be included in the partial interpretation as well - if they have not been included already. At the same time other meanings of those words will be removed from the CHUNK. If a meaning ends up with no words realising it, all its incident edges will be removed from the CHUNK. Once there are no more edges in the CHUNK the partial interpretation is done. All that remains are words that could not be disambiguated because nothing connected them to the context.

I am ready to present the nondeterministic algorithm for finding partial interpretations of S. This will allow me to describe the algorithm in a clear way and provide, the best way of understanding how the algorithm works. It will then be fairly straightforward to modify the algorithm to determinancy.

Output consist of a set of lexemes, $I^P L$, and a set of relations between them, $I^P R$. Figure 28 shows all the possible partial interpretations of S_1 , even though the algorithm as it is, only finds one.

Algorithm: $I^P(S)$ – nondeterm

0. read $S \leftarrow "s_1, \dots, s_n"$;
1. with S build CHUNK C(RVOCAB, LINKS);
(incl. M_0, W_0, R_0 and related subsets)
2. $I^P L \leftarrow \emptyset$; $I^P R \leftarrow \emptyset$;
(current partial interpretation, i.e., Lexemes and Relations)
3. WHILE nonempty(LINKS) do
4. choose and delete from LINKS an edge $Rel(m_i, m_j)$;
5. IF nonempty($W_0(m_i)$) AND nonempty($W_0(m_j)$) THEN
6. add $Rel(m_i, m_j)$ to $I^P R$;
7. FOR $w \in W_0(m_i)$ do add (w_i, m_i) to $I^P L$;
8. FOR $w \in W_0(m_j)$ do add (w_j, m_j) to $I^P L$;
9. FOR $m \in M_0$ do $W_0(m) := W_0(m) \setminus \{w_i, w_j\}$;
10. FOR $w \in W_0$ do $M_0(w) := M_0(w) \setminus \{m_i, m_j\}$;
11. report $(I^P L, I^P R)$.

The WHILE-loop in line 3 operates on a copy of the LINKS part of the CHUNK as already described, recursively choosing and deleting from it possible relationships between the nouns of the sequence.

6.5.2 From non-determinism to determinism

Solely the "choosing" in line 4 causes the nondeterminism of the above algorithm. In the trace of Figure 27 (- different solutions shown in Figure 28), the edge that is chosen first, enters the partial interpretation and the other one does not, since inclusion of either depends on a different meaning of w_3 . To achieve determinism, backtracking with regard to choice of edge in each execution of Line 4 is necessary. Backtracking is easy in Prolog, which is the language I will use for my prototypes. The clarity of the algorithm would suffer unnecessarily if I were to illustrate backtracking explicitly, so I refrain from doing so. Instead the non-deterministic algorithm adequately illustrates the cycle that finds each of the solutions.

Let us examine a cycle of the S_1 sequence, tracking the store line for line:

Below, the "Store" represents various instantiations of variables after execution of the associated pseudo-instruction. The notation is as follows : <ul style="list-style-type: none"> - Store \leftarrow [instantiations] - within the brackets are all current instantiations. - Store \leftarrow Store[changes] - Store is altered according to the changes in the brackets, otherwise it remains as before executing the associated pseudo-instruction. - Store \leftarrow Store \cup [additions] - Store is augmented by instantiations represented by the additions in the brackets. 		
Line	Pseudo-instruction	Store after execution / Comments
0	read S \leftarrow "s ₁ , ..., s _n ";	Store \leftarrow S=["w ₁ , w ₂ , w ₃ "].
1	with S build CHUNK C(RVOCAB, LINKS); (incl. M ₀ , W ₀ , R ₀ and related subsets)	Store \leftarrow Store \cup [RVOCAB = { (w ₁ , M ₀ (w ₁)), (w ₂ , M ₀ (w ₂)), (w ₃ , M ₀ (w ₃))} M ₀ (w ₁) = {m ₁ , m ₂ } M ₀ (w ₂) = {m ₁ , m ₃ } M ₀ (w ₃) = {m ₄ , m ₅ } LINKS = R ₀ R ₀ = {hyp(m ₁ , m ₄), hyp(m ₃ , m ₅)} W ₀ = {w ₁ , w ₂ , w ₃ } M ₀ = {m ₁ , m ₂ , m ₃ , m ₄ , m ₅ } W ₀ (m ₁) = {w ₁ , w ₂ } W ₀ (m ₂) = {w ₁ } W ₀ (m ₃) = {w ₂ } W ₀ (m ₄) = {w ₃ } W ₀ (m ₅) = {w ₃ }]
2	I ^P L \leftarrow \emptyset ; I ^P R \leftarrow \emptyset ;	Store \leftarrow Store \cup [I ^P L= \emptyset , I ^P R= \emptyset].
3	WHILE nonempty(LINKS) do	
4	choose and delete from LINKS an Edge: hyp(m ₁ , m ₄);	Store \leftarrow Store[R ₀ = {hyp(m ₃ , m ₅)}] \cup [Edge=hyp(m ₁ , m ₄)]
5	- IF nonempty(W ₀ (m ₁)) AND nonempty(W ₀ (m ₄)) THEN	/*Make sure ending points of the edge can be realised*/
6	- - add h(m ₁ , m ₄) to I ^P R	Store \leftarrow Store[I ^P L= \emptyset I ^P R={hyp(m ₁ , m ₄)}]

7	-- FOR $w_i \in W_0(m_i)$ do add (w_i, m_i) to $I^P L$	Store ₅ ← Store[$I^P L = \{(w_1, m_1), (w_2, m_1)\}$ $I^P R = \{\text{hyp}(m_1, m_4)\}$]
8	-- FOR $w_j \in W_0(m_j)$ do add (w_j, m_j) to $I^P L$	Store = Store[$I^P L = \{(w_1, m_1), (w_2, m_1), (w_3, m_4)\}$ $I^P R = \{\text{hyp}(m_1, m_4)\}$]
9	-- FOR $m \in M_0$ do $W_0(m) := W_0(m) \setminus \{w_i, w_j\}$;	Store ← Store[$W_0(m_1) = \emptyset$, $W_0(m_2) = \emptyset$, $W_0(m_3) = \emptyset$, $W_0(m_4) = \emptyset$, $W_0(m_5) = \emptyset$]
10	-- FOR $w \in W_0$ do $M_0(w) := M_0(w) \setminus \{m_i, m_j\}$;	Store ← Store[$M_0(w_1) = \{m_2\}$, $M_0(w_2) = \{m_3\}$, $M_0(w_3) = \emptyset$]
After the first loop the algorithm re-enters in line 3, and since there are still edges in LINKS it succeeds in completing the loop. The only change made to the store is removal of final edge in Line 4. The condition in Line 5 is not met since all $W_0(m)$ sets are now empty. When the WHILE loop is tried a third time, there are no more edges in LINKS and the algorithm terminates.		
11	visualise I^P ;	Store ← [S = "w ₁ w ₂ w ₃ " RVOCAB = { (w ₁ , M ₀ (w ₁)), (w ₂ , M ₀ (w ₂)), (w ₃ , M ₀ (w ₃))} M ₀ (w ₁) = {m ₂ } M ₀ (w ₂) = {m ₃ } M ₀ (w ₃) = ∅ LINKS = R ₀ = ∅ W ₀ = {w ₁ , w ₂ , w ₃ } M ₀ = {m ₁ , m ₂ , m ₃ , m ₄ , m ₅ } W ₀ (m ₁) = ∅ W ₀ (m ₂) = ∅ W ₀ (m ₃) = ∅ W ₀ (m ₄) = ∅ W ₀ (m ₅) = ∅ $I^P L = \{(w_1, m_1), (w_2, m_1), (w_3, m_4)\}$ $I^P R = \{\text{hyp}(m_1, m_4)\}$]

Figure 27: Trace of one non-deterministic cycle of the algorithm.

6.5.3 How partial interpretations relate to complete interpretations

The Partial interpretation found is the one shown as $I^P_1(C_1)$ in Figure 28. Incidentally, it turns out to be a complete interpretation, since it assigns exactly one meaning to all the words in W_0 . However interesting, this is not the general case. Had the edge, $\text{hyp}(m_3, m_5)$, been chosen first the resulting interpretation of the algorithm would have been the incomplete one shown as $I^P_2(C_1)$ in Figure 28.

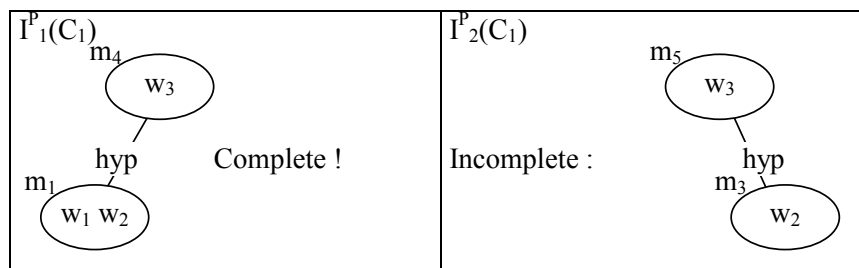


Figure 28: The two possible partial interpretations of C_1 .

Furthermore when regarding the partial interpretations of Figure 27, it becomes clear that they are each representatives of one or more of the complete interpretations of C_1 shown in Figure 26:

$I^P_1(C_1)$ represents those complete interpretations that involve the edge $h(m_1, m_4)$, i.e., $I_1(C_1)$, $I_3(C_1)$ and $I_7(C_1)$.

$I^P_2(C_1)$ represents those complete interpretations that involve the edge $h(m_3, m_5)$, i.e., $I_4(C_1)$ and $I_6(C_1)$.

When looking closely at the available values of the system upon completion of the partial interpretation cycle, especially the configuring of the $M_0(w)$ sets, it should be clear that all complete interpretations corresponding to that particular partial interpretation can be generated as follows :

Given partial interpretation $I^P_1(C_1)$ and the following:

$M_0(w_1) = \{m_2\}$

$M_0(w_2) = \{m_3\}$

$M_0(w_3) = \emptyset$,

- Moving w_1 to any of the meanings in $M_0(w_1) = \{m_2\}$ will result in a complete interpretation. : $I_1(C_1)$. Since all original concepts are still realised, no harm is done.
- Moving w_2 to any of the meanings in $M_0(w_2) = \{m_3\}$ will result in a complete interpretation. : $I_3(C_1)$. Since all original concepts are still realised, no harm is done.
- Moving w_3 will result its original concept to become unrealised, and thus is not allowed.

Similarly:

Given partial interpretation $I^P_2(C_1)$ and

$M_0(w_1) = \{m_1, m_2\}$

$M_0(w_2) = \{m_1\}$

$M_0(w_3) = \emptyset$,

- Moving w_1 to any of the meanings in $M_0(w_1) = \{m_1, m_2\}$ will result in a complete interpretation. : $I_4(C_1)$, $I_6(C_1)$. Since all original concepts are still realised, no harm is done.
- Moving w_2 will result its original concepts to become unrealised, and thus is not allowed.
- Moving w_3 will result its original concepts to become unrealised, and thus is not allowed.

This indicates that a procedure can be formulated that generates any complete interpretation that is represented by a given partial one, moving members between sets, maintaining that the original sets of the partial interpretation never becomes empty. Note that only those complete

interpretations that consist of only *isolated* lexemes cannot be reached this way. I will not, however, develop such a procedure in the present project.

6.5.4 How to compare interpretations

My approach to measuring “goodness” leans heavily on Grice’s Cooperative Principle, (Grice, H. P. 1975), that is - the author can be assumed to have gone to considerable length to make the text as clear as possible. In doing so he or she will - on purpose or unwittingly - have left cooperative indicators behind in the text as to the meaning that he/she intended for it to convey. My goodness metric will involve those cooperative indicators that can be accessed by the system. Many such indicators can be identified. The ones I chose for analysis include:

Cooperative Indicators

- Degree of semantic coherence : This is the single most important indicator in my analysis. If readings of two nouns , i.e., their meanings, “match” each other via lexical relations, an interpretation involving those readings are be considered better than those that do not. As an extension : The fewer unrelated lexemes involved in an interpretation the higher that interpretation should score. It is important to point out that the comparison can involve only those relationships that is represented in the MRD.
- Degree of disambiguation : The larger the portion of nouns that are disambiguated by this partial interpretation the better it will considered to be. Here I consider both the ratio between in the number of nouns in W_c relative to the number of nouns in W_0 as well as the ratio between the number of instances represented by the two W -sets respectively. That is n instances are represented by W_0 (the length of the original input sequence), whereas some number $n_c \leq n$ of instances are represented by W_c .
- Frequency of nouns : The nouns in W_0 that the more instances in S should receive more attention than nouns with fewer instances in S , since they are possibly more important to the core of the content of the text.
- Sequential distance between instances : This cooperative indicator maybe best illustrated by an abstract example depicted in Figure 29. Consider a sequence : ” $s_1 \dots s_4 \dots s_{11} \dots$ ” of different nouns. Suppose that there is a choice between lexemes to be represented by the instance s_1 : one that connects to the context via the instance s_4 and another that does so via the instance s_{11} . This cooperative indicator requires explicit tracking of sequential positions of instances and dictates that the alternative, involving those instances that are closest to each other in the sequence, is to be chosen. Here the lexeme for s_1 that connects with the instance s_4 is chosen. To see this, simply regard the indices of the instances as their relative positions in the S , compare their differences and choose the smaller (Here, $4-1=3$ vs. $11-1=10$). If the alternatives are of equal distance to

the context, no decision can be made on the grounds of sequential distance between instances.

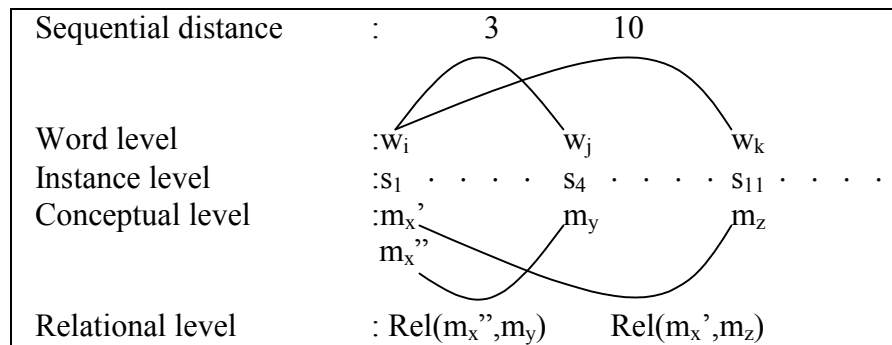


Figure 29: The sequential distance between instances has a role in disambiguation.

In other words, a better interpretation will hopefully satisfy the Cooperative Principle (as I have interpreted its implications in the bulleted list above, that may very well be refined and expanded) to a higher degree than infelicitous interpretations.

6.5.5 Two scoring schemes

Which interpretations are “good”? This is where the cooperative indicators inspired from Grice’s Cooperative Principle come in. As the grounding of these I place the lexical relationships as they can be recognised between the possible lexemes of the CHUNK, i.e., before any interpretation is actually done.

Regarding first the complete interpretations of Figure 26, I will present two scoring schemes. The first - rather detailed one - will attempt to explicitly formalise the cooperative indicators identified and described in the previous section. (I will later apply the scoring to the partial interpretations and compare the results.)

The second scoring scheme is a much simpler scheme that never the less serves as a less explicit formalisation of the cooperative indicators. In essence it relies on only the first of the cooperative indicators, *Degree of semantic coherence*. The main tendencies of both scoring schemes are very comparable, and the differences between them mainly concern their respective degrees of granularity.

Scoring Scheme1: Formalising the Cooperative indicators

As an example of a rather detailed scoring system, I will, for each indicator, award points and penalties to the individual complete interpretations of $S_1, I_{1-8}(C_1)$, 1-8 for short, from Figure 26, according to how “good” they are deemed by that indicator:

- +1 point for “good”
- no points for indifferent
- 1 point for “not good”

Degree of semantic coherence: Minimizing the number of unrelated lexemes will rank the eight interpretations as follows :

- +1, no isolated lexemes : 7
- 1, all isolated lexemes : 2 and 5

Degree of disambiguation: Well, these are complete interpretations of the sequence. All the words in the sequence are assigned a meaning in each of the interpretations. This indicator doesn't yield anything new in this case. Awarded points :

- +1 : 1, 2, 3, 4, 5, 6, 7 and 8

Relative order of nouns: None of the explicit edges can be placed before the last word in the sequence, w_3 , is met since that is the only word that can realize any of the meanings m_4, m_5 , without which both edges would lack an ending point. However, this indicator should prefer those interpretations that have w_1 and w_2 as synonyms and points are awarded accordingly:

- +1 : 7 and 8

Frequency of nouns : Again we are dealing with a sequence of three different nouns. Each occurs exactly one time in the sequence. Since none is surpassed they all score maximally:

- +1 : 1, 2, 3, 4, 5, 6, 7 and 8

Sequential distance of instances : The closest pairs of instances in the sequence are (s_1, s_2) and (s_2, s_3) , represented by nouns w_1, w_2 and w_3 respectively. Any interpretation that involves the connection of both of these pairs will be awarded 1 point and those that connect neither of them are penalized by one point. Points are awarded follows:

- +1 pts, for both : 7
- 1 pts, for none : 2 and 5

The result is:

Scoring Scheme 1								
Complete Interpretation	1	2	3	4	5	6	7	8
Total points	+2	0	+2	+2	0	+2	+5	+3

The tendency should be clear. According to the cooperative indicators, 7 is the winner. The runner-up is 8. Contestants 1,3,4 and 6 fought well while contestants 2 and 5 should not give up their day job.

Scoring Scheme 2: Connected lexemes

An alternative scoring scheme, the one I use in the actual prototype, makes use of the notion of “connected lexemes”. A lexeme is connected in an interpretation if it is member of the ending point of at least one explicit edge in that interpretation. Thus the number of connected lexemes in an interpretation correlates closely to the notion of semantic coherence and

thus corresponds nicely with the first cooperative indicators in my list, *Degree of semantic coherence*.

An interpretation scores +1 point for each connected lexeme it involves compared to the maximum number of lexemes in a complete interpretation. (Since synonyms are not represented by explicit edges in the present version CHUNK, they will have to be dealt with in a refinement of the scoring system.)

The scores of the eight complete interpretations according to this second scoring scheme are:

Scoring Scheme 2								
Complete Interpretation	1	2	3	4	5	6	7	8
Total points	2/3	0/3	2/3	2/3	0/3	2/3	3/3	0/3

Comparing the Schemes

Obviously, the extreme simplicity of scoring scheme 2 is very appealing because it can be applied considering the number of connected lexemes of the interpretation. To see that the two scoring schemes are different expressions of the same tendency let us convert their respective results to percentages of maximum score. We get the following comparison:

Comparison								
Complete Interpretation	1	2	3	4	5	6	7	8
Scoring Scheme 1	40%	0%	40%	40%	0%	40%	100%	60%
Scoring Scheme 2	67%	0%	67%	67%	0%	67%	100%	0%

It is clear that scoring scheme 2 pertains the tendency from scoring scheme 1, except in the case of interpretation 8 due to the current inability of scoring scheme 2 to recognize synonyms properly. Synonyms should be dealt with as a refinement. The scorings of the partial interpretations from Figure 28 are as follows. Note that computing the scores of partial interpretations according to scoring scheme 2 becomes extremely easy. The score of each reflects the length of its $I^P L$.

Comparison		
Partial Interpretation	1	2
Related Complete Interpretations	1,3,7	4,6
Scoring	100%	40%

Scheme 1		
Scoring Scheme 2	100%	67%

Thus the two scoring schemes can be seen as two opposing extremes on an axis of complexity vs. simplicity. It is possible to compare the interpretations in several different ways, and it remains a task to experiment with different scoring schemes to find a balance between, degree of granularity one side and complexity of computation on the other.

6.6 Remarks

The provided solution, i.e., the three tools : CHUNK, partial interpretation algorithm and scoring scheme 2, constitute a solid foundation for the necessary experiments to be carried out. It is important, however to point out that several issues still need to be addressed for the results to properly reflect the quality of the solution.

6.6.1 Relations between lexemes

The solution handles relations between meanings as I think they should be treated. Relations between lexemes, i.e., antonymy and synonymy, still need attention.

When an antonymous relationship is picked out in Line 4 of the algorithm, the algorithm should add the edge between the associated meanings as usual. It should however only add those lexemes that actually occur as the arguments of the relationship – not possible synonyms. That is, if the chosen edge is an antonymous one, e.g., $\text{ant}(w_1, m_1, w_3, m_4)$, it will

	Store in Line 4	Resulting additions to I^P
Hyponymous or meronymous relationship - add edge and <u>all</u> lexemes involving the respective arguments	[... $R_0 = \{\text{hyp}(m_1, m_4), \dots\}$ $W_0(m_1) = \{w_1, w_2\}$ $W_0(m_4) = \{w_3\}$...]	$I^P R \leftarrow \{\text{hyp}(m_1, m_4)\}$ $I^P L = \{(w_1, m_1), (w_2, m_1), (w_3, m_4)\}$
Antonymous relationship Add only antonymous lexemes	[... $R_0 = \{\text{ant}(w_1, m_1, w_3, m_4), \dots\}$ $W_0(m_1) = \{w_1, w_2\}$ $W_0(m_4) = \{w_3\}$...]	$I^P R \leftarrow \{\text{ant}(m_1, m_4)\}$ $I^P L = \{(w_1, m_1), (w_3, m_4)\}$
Synonymous relationship Add all synonymous lexemes	[... $R_0 = \{\text{syn}(w_1, m_1, w_2, m_1), \dots\}$ $W_0(m_1) = \{w_1, w_2, w_3\}$ $W_0(m_4) = \{w_3\}$...]	$I^P R \leftarrow \emptyset$ $I^P L = \{(w_1, m_1), (w_2, m_1), (w_3, m_1)\}$

Figure 30: Examples of the three kinds of relationships and their proper treatment.

dictate that the two lexemes be added, i.e., (w_1, m_1) and (w_3, m_4) . Lines 7 and 8 of the algorithm, should add only those. Likewise only the associated nouns should be removed from $W_0(m_1)$ and $W_0(m_4)$ respectively.

With regard to synonyms, they are already included as a bonus when their associated meaning is otherwise connected in the partial interpretation. They should however also be included in the partial interpretation when at least two of the nouns realising the meaning has not been assigned to other meanings. The proper way to remedy this is to generate explicit synonymous edges in LINKS when building the CHUNK, e.g., $\text{syn}(w_1, m_1, w_2, m_1)$. That is, an edge of $\text{hyp}(m_i, m_j)$ is included in LINKS then it is discovered that $h(m_i, m_j)$ is an entry in the hyponym database of the MRD. Similarly an edge, $\text{syn}((w_i, m), (w_j, m))$, should be included in LINKS when it is discovered that both entries $s(m, w_i, _)$ and $s(m, w_j, _)$ are in the synset database of the MRD. When such an edge is chosen in Line 4, the algorithm should check that at least two nouns realising the associated meaning is still unassigned - if so, it should add all the unassigned nouns capable of realising that meaning to the partial interpretation. Note that while this will allow the algorithm to capture synonyms, the associated words are prevented from being assigned other meanings in that interpretation. Backtracking will secure that all possible partial interpretations are still reached.

All in all this indicates a separate version of lines 5-10 of the WHILE-loop for each of the three kinds of relationships in LINKS. Also the edges corresponding to the bonus synonyms, gained from hyponymy and meronymy, should be removed from LINKS.

These parts of the algorithm have not all been implemented yet but a sketch of the needed modifications to the algorithm is as follows - assuming that the CHUNK-generating procedure has been modified to include synonymy edges in LINKS. Note that these changes involve some abstractions over the arguments of the different kinds of relationships.

<p>Algorithm: $P(S)$ <i>with modified CHUNK</i>, backtracks. Not implemented.</p> <p>0. read $S \leftarrow "s_1, \dots, s_n"$;</p> <p>1. with S build CHUNK $C(\text{RVOCAB}, \text{LINKS})$; (incl. M_0, W_0, R_0 and related subsets <i>and synonymous relationships in LINKS</i>)</p> <p>2. $I^P L \leftarrow \emptyset$; $I^P R \leftarrow \emptyset$; (current partial interpretation, i.e., Lexemes and Relations)</p> <p>3. WHILE nonempty(LINKS) do</p> <p>4. choose and delete from LINKS an edge $\text{Rel}(\text{Arg1}, \text{Arg2})$;</p> <p>5. IF edge = $\text{syn}((w_i, m_k), (w_j, m_k))$ THEN</p> <p>6. IF $\text{length}(W_0(m_k)) > 1$ THEN</p> <p>7. FOR $w \in W_0(m_k)$ do /* all available synonyms */</p> <p>8. add (w, m_k) to $I^P L$;</p> <p>9. $M_0(w) \leftarrow M_0(w) \setminus \{m_k\}$;</p> <p>10. FOR $m \in M_0$ do $W_0(m) := W_0(m) \setminus \{w\}$;</p>
--

```

11.   remove from LINKS all edges syn(.,mi,.,mi);
12.   ELSE IF edge = ant((wi,mi),(wj,mj)) THEN
13.     IF nonempty(W0(mi)) AND nonempty(W0(mj)) THEN
14.       add(ant(mi,mj) to IPR;
15.       add (wi,mi),( wj,mj) to IPL;    /* only antonyms */
16.       FOR m∈M0 do W0(m):= W0(m)\{wi,wj};
17.       FOR w∈W0 do M0(w):= M0(w)\{mi,mj};
                                           (continued ...)

18.   ELSE IF nonempty(W0(mi)) AND nonempty(W0(mj)) THEN
19.     add Rel(mi,mj) to IPR
20.     FOR w∈W0(mi) do add (wi,mi) to IPL;
21.     FOR w∈W0(mj) do add (wj,mj) to IPL;
22.     FOR m∈M0 do W0(m):= W0(m)\{wi,wj};
23.     FOR w∈W0 do M0(w):= M0(w)\{mi,mj};
24.     remove from LINKS all edges syn(.,mi,.,mi) or syn(.,mj,.,mj);
25.   report (IPL, IPR).

```

Figure 31: Different actions have to be taken depending on the kind of lexical relationship in question - synonymy, antonymy or hyponymy/meronymy respectively.

6.6.2 Complexity analysis

I have already pointed out, that the complexity of complete interpretation is upward bounded by $O(K^n)$.

With regard to partial interpretations it should be clear that that the complexity of each of the FOR loops are of an complexity proportional to n . That is either the number of elements in W_0 , which is at most n , or the number of members of M_0 , which is at most Kn . (K is the maximum number of meanings of any noun in WordNet). The while loop is entered once for each of the edges in R_0 . When going from non-determinism to determinism the while-loop is entered $|R_0|!$ times. So analysing the complexity of the partial interpretation-algorithm must involve analysing on the number of members in R_0 and thereby the nature of relations involved:

Properties of the lexical relations of ontologies usually include:

A: Relations like hyponymy and meronymy are basically treelike structures with the odd multiple inheritance now and then, (formally the structure is called a DAG, short for Directed Acyclic Graph. The “directedness” is eliminated from the complexity in this case).

B: It is safe to assume that at most one relationship can exist between any given pair of lexemes. That is, if there is a hyponymous relationship between concepts m_i and m_j , there will never be instances of meronymy between those concepts. Neither will there exist antonymous lexemes (w_i, m_i) , (w_j, m_j) . Of course synonymy between such lexemes cannot exist either.

I use undirected graphs to represent both the CHUNK and solutions. The maximum number of edges in an undirected graph of Kn vertices is $(Kn^2 - Kn)/2$. Further more I must address the fact that I use three different relationships, and four when counting synonyms, this should give us an upper limit of members of R_0 of $(4Kn^2 - 4Kn)/2$. This measure is however not quite fair for the following reasons:

- 1) The edges in R_0 must reflect the underlying Ontology in the MRD. Ontologies are graphs rather than trees, since they do involve multiple inheritances. Multiple inheritances in ontologies are quite rare, however. Because of property A, above, it is a much more fitting description of each of the conceptual relationships, to say that their edge/vertex ratio is much closer to $Kn-K$.

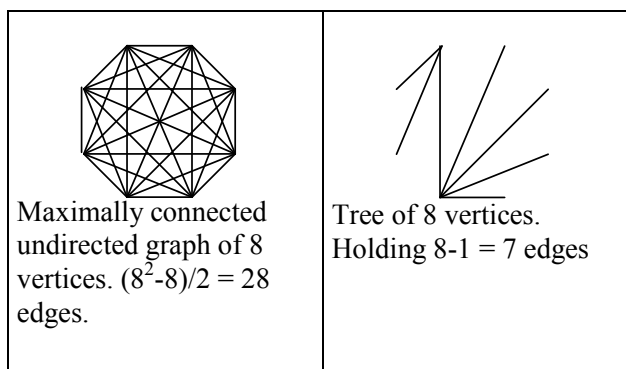


Figure 32: The ratio between vertices and edges in a network representing an lexical relation between concepts is much closer to that of trees, rather than that of maximally connected undirected graphs.

- 2) If we assume a treelike structure for each of the four relations we choose to include - each of $Kn-K$ edges. Joining those structures together in one graph accordingly is at most the sum of the edges, i.e., $4Kn-4K$. However as an edge-limiting consequence of property B above, there will be no instances of parallel edges in the junction of the trees.

These observations should demonstrate that the number of edges in R_0 is much closer to $4Kn-4K$ than to $(4Kn^2 - 3Kn)/2$. As a pragmatic confirmation of the properties, A and B, let us compare to the statistics of the WordNet databases.

The documentation of the WordNet version, v1.6, that I used for this project states that the databases involve :

$W = 57000$ distinct nouns eligible for membership of W_0 , organised into:

$M = 48.800$ distinct synsets eligible for membership of M_0 .

Inspection of the individual databases reveals that there are:

- Approximately 118000 distinct lexemes.

- Approximately 68000 pairs of nominal synsets in hyponymy to each other. (I.e., about $1.4|M|$)
- Approximately 19400 pairs of nominal synsets in various kinds of meronymy to each other. (I.e., about $0.4|M|$)
- Approximately 7800 pairs of nominal lexemes in antonymy to each other.
- Synonymy is a looser estimate, but we have on average 2.4 lexemes for each meaning, giving us an low estimate of 1.2 synonymous pair for each distinct meaning on average, i.e., approximately 57600 pairs of synonymous lexemes is not to far off.

In total approximately:

$R = 152800$ relationships eligible for membership of R_0 .

First of all, 152800 edges is a far cry from $4|M|^2$. In fact it just about $3.1|M|$, corresponding to $3.1Kn$.

Secondly, we can identify that there are about 2.1 lexemes/noun. This tells us that each noun has on average 2.1 different meanings. Should a text involve exactly one instance of all 57000 nouns in W we would have to generate and compare about 2.1^{57000} different complete interpretations, corresponding to K^n . That is indeed a very large number:

3,1608205002210808705367585146e+18366

As a further limitation on the resulting complexity of the algorithm is that the algorithm makes shortcuts - including synonymous edges from links en bloc - whenever possible (about 38% of the edges in R_0 involve synonymy). As a consequence the WHILE loop will never be invoked $|R_0|!$ times - the factorial (!) was introduced as a consequence of backtracking in the deterministic version of the algorithm. Rather R_0 will in the vast majority of cases become empty shortly after each edge has been chosen first once. That leads to an estimate of $|R_0|^x$ complete executions of the WHILE loop on average, where the exponent x corresponds to the number of times the WHILE LOOP is on average invoked. My argument is that x will always be a small fraction of $|R_0|$. (This estimate is admittedly vague but proving a further restriction on this exponent $x < |R_0|$ is beyond the scope of this thesis, and will not be attempted.)

As a conclusion it can be assumed – even if not formally proved - that while $O(n!)$ remains the worst case complexity - and this is not much better than the $O(K^n)$ of the complete interpretation – the vast majority of input sequences will be done much easier, in time proportional to n^x . Experimenting with the prototype will perhaps shed some light on the magnitude of this X .

Finally, it is obvious that working out a good scoring scheme is of paramount importance to the quality of the solutions. Scoring scheme 2 will form the basis for the experiments to be carried out with the prototype and the experimental corpus. The experiments should however, explore the possibilities for further refinement to the scoring scheme.

I will introduce one minute modification to scoring scheme 2. One lexeme may be connected to each of the other lexemes in a given the interpretation. The contribution to the score of an interpretation that the inclusion of a given lexeme entails, should reflect the “cardinality” of that lexeme in that interpretation. Thus each connected lexeme of an interpretation will make its contribution to the score every time it is the ending point of an edge in the graph of that interpretation. So if a particular vertex has more than one edge going to or from it (see vertex A in the left graph of Figure 33), it contributes to the score of all of those edges. This modification makes the scoring scheme prefer a coherent graph to one consisting of several distinct sub-graphs, also called components in graph theory.

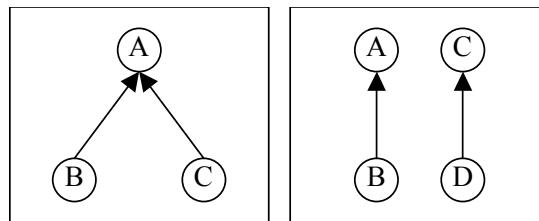


Figure 33: In both the above interpretations above we have two edges in total. In the one to the left we see two edges connecting the A vertex in one coherent component of three vertices. The interpretation to the right involves two distinct components of two vertices each. The scores of the edges depend on the scores of their ending points. Say for instance that each vertex A,B,C and D above realise 1 instance in the sequence. The graph to the left would originally only score 3 whereas the one to the right would score 4. The modification causes the A vertex to score twice in the left interpretation above since it is an ending point of two edges. Thus the modified scoring system would make both alternatives score the same namely 4.

7 Output – Examples and Experiments

In this chapter I am going to put my prototype program to the test of parts of the CIVIII corpus as well as other pieces of natural language data. I will quickly sum up the decisions I have made.

7.1 A re-introduction

The program works as a **skimming** function that takes English text as argument and produces a series of “good” interpreting graphs for the nouns of that text and reports the “best” one(s) of these as output.

- The text is **skimmed** taking into regard only the sequence of **nouns** as produced by the Tosca-ICLE tagger applied to the original (though slightly pre-processed) text. The sequence also contains information about paragraph boundaries in the original text. (In effect the noun sequences are produced by hand. I have, however, shown that a fairly straightforward rule governed algorithm can produce the sequence from the tagger-output.)
- Graphs are defined as $G(V,E)$ (i.e.: a set of vertices and a set of edges). I have chosen to let the vertices represent the lexemes of the “interpretation” leaving the edges to represent lexical relationships between the lexemes.
- The nouns are looked up in the MRD and the interrelationships between their possible meanings are used as a guide to the interpreting graphs that best satisfy the measure on “goodness”.
- The measure of “goodness” is taken care of by the scoring scheme developed in the previous chapter. The scoring of a particular graph, representing one particular set of incomplete interpretations of the sequence, depends on the number of nouns that are represented by the connected lexemes in the interpreting graph. The more instances that are disambiguated this way the better the interpretation. Also if the same lexeme is part of several relationships it will score once for each such relationship.⁷
- The interpretations are partial since nouns that do not have connected meanings in the graph are not disambiguated, thus choosing any of their meanings will not affect the score.

When illustrating the many graphs of this chapter, I have for reasons of clarity, incorporated a commercially licensed graph drawing application

⁷ In effect the score of an included edge is the sum of the scores of its ending points. Each ending point represents a sense and the words realizing that sense. Each instance in the sequence represented by a word in a vertex add 1 to the score of the vertex. Consider some edge involving two ending points. Suppose one of these ending points represents two instances in the sequence that are assigned the sense associated with the vertex, while the other vertex represents only one instance. The inclusion of the edge in an interpreting graph will add 3 to the score of that graph. Also a given vertex will add its score to the score of the interpreting graph every time it is an ending point of an edge of that graph.

called *Graphlet*. *Graphlet* was developed at University of Passau and is distributed by BRAINSYS Informatik Systeme GmbH. I was kindly allowed by Brainsys to use the system free of charge for the purpose of this project. For full information on *Graphlet* and the *Graph Meta Language(GML)* that it employs, please see the web page of the distributor, (<http://www.brainsys.de>).

Graphlet is a quite sophisticated application that addresses the problem of arranging an arbitrary graph so that it fits on a sheet of paper or a certain portion of it (- a computability problem in its own right). I am not nearly employing all of its tools, since I merely use it to draw the graphs for me in a portable format. It requires me to output the various graphs in the *Graph Meta Language (GML)*, which is more or less the theoretical format for graphs as most textbooks have it. The resulting GML file can then be manipulated in the *Graphlet* editor application and exported to the postscript-printing format. This means that the graphs can be drawn, printed and imported in a presentable manner using various layout algorithms.

7.2 A simple example

I will now return to the simple example sequence mentioned in the introduction of this paper, namely: “*society civilization culture*“. It is assumed that the sequence represents the nouns in some natural language text. Along the way I will point out the various output files and explain the information they hold.

7.2.1 Interpreting graph for “*society civilization culture*“

The sequence itself will be represented in data-file “SimpleSequence.txt” as follows:

```
p.
society.
civilization.
culture.
```

Figure 34: Contents of the file “SimpleSequence.txt”. Each noun in the original text is represented in the sequence once for each time it occurs. “p” is a marker representing the beginning of a new paragraph in the original text. Each item occupies its own line and each line is terminated by a period, “.”. Thus each item conforms to the syntax of a PROLOG fact.

A complete interpretation of this sequence must involve exactly three lexemes so we try to find unambiguous meanings for each of the three nouns. Recall that a lexeme in my definition is a tuple of the format: (word,meaning). So what the system will try to establish is an unambiguous meaning for each of the words in the sequence, using lexical relationships to decide proper meanings.

Looking up *society* in the MRD reveals that has the following possible meanings:

<p>s1 : (the state of being with someone; "he missed their company"; "he enjoyed the society of his friends")</p> <p>s2 : (the fashionable elite)</p> <p>s3 : (a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today")</p> <p>s4 : (an extended social group having a distinctive cultural and economic organization)</p>
--

Figure 35: *Society* - the different meanings of the polysemous noun. “s1-s4” each represents their own 9-digit SynSet ID, which I abbreviated for reasons of simplicity. The descriptive text is not part of the senses themselves but rather their corresponding explanatory wording in the glossary database of WordNet and provided here as a service to the reader. The “glosses” are not used in any way in the algorithm.

Since *society* is the first word in the sequence and lexical relationships requiring two meanings or lexemes, lexical relationships cannot help us disambiguate *society* in isolation. We are for now stuck with 4 different interpretations of the sequence so far.

Reading the next noun reveals that *civilization* can only mean the following⁸:

<p>s5 : (a society in an advanced state of development)</p>
--

Figure 36: *civilization*. This noun is monosemous or unambiguous in the MRD, having only one possible meaning shown here as “s5”.

When checking what possible relationships may be at play the MRD reveals that indeed there is a hyponym relation between s5 and s4. This means that if the two words are assigned these meaning it makes sense to state that “*a civilization is a society*”. Among the different possible interpretations of the two words this is the only one that relates the resulting lexemes to each other semantically (as far as the information in the MRD goes anyway). The interpreting graph so far has:

$G(V,E)$:

$V = \{(society,s4),(civilization,s5)\}$ and

$E = \{hyp(s5,s4)\}$.

This graph involves two connected lexemes namely:

(society,s4)

(civilization,s5)

⁸ Because the noun *civilization* has only one possible meaning in the MRD it is in essence already disambiguated. However since my system requires participation in a lexical relationship in order for a noun to be disambiguated, *civilization* will be counted as not disambiguated unless it’s meaning is in fact related to other meanings of a particular interpreting graph. This is a quirk of course, but one I decided not to remedy since my focus is on lexical relationships and what can be gained from them. So I choose to regard all nouns as polysemous until proved otherwise by lexical relation participation. Finally note that in fact *civilization* is used in two different ways in the corpus, sometimes referring to the game and sometimes to the usual social meaning of the word. Only the latter is ever recognised in the experiments. Since the former is particular to the context of the PC game and invented for it, it does not figure in the MRD.

Each of these occurs only once in the sequence. Thus the scoring of this interpretation is 2, compared to 0 for any other interpretation of the two nouns.

Reading the final word in the sequence we get that the noun *culture* has the following set of possible meanings:

s6 : (a particular civilization at a particular stage)
s7 : (all the knowledge and values shared by a society)
s8 : (the tastes in art and manners that are favoured by a social group)
s9 : ((biology) the growing of microorganisms in a nutrient medium (such as gelatin or agar); "the culture of cells in a Petri dish")
s10 : (the raising of plants or animals: "the culture of oysters")

Figure 37: *culture*: Another polysemous noun.

Checking for relationships we find that we have a hyponym relation between s6 and s5 stating that “*a culture is a civilization*” when the words have the proper meanings – s6 and s5 respectively. So an interpreting graph involving both relationships will be the preferred reading of the complete sequence. That graph can be represented like this:

$G(V,E)$, where

$V = \{(society,s4),(civilization,s5),(culture,s6)\}$ and

$E = \{hyp(s5,s4),hyp(s6,s5)\}$

This graph contains three connected lexemes and the normal score would be 3. However, since s5 is involved in two separate relationships it scores twice. The total score of this graph is then 4, easily beating any interpretation not employing the lexemes (culture,s6) or (society,s4) that would each score at most 2.

Observe that there are in total 20 different interpretations of these three words as it can be seen by the product of the respective numbers of readings of the words, $4 \times 1 \times 5$. These different interpretations are illustrated in table of Figure 38. See how the scoring scheme singles one of them out. It does so by considering only the two relational edges and disregarding the unrelated senses altogether instead of comparing all twenty possible solutions. This could just be what a human reader might do, i.e.: *turn to the context for clarification* in the interpreting of the words.

Using *Graphlet* the chosen graph can be presented as the GML-file shown in Figure 39. The relationships between the three lexemes are easily seen as the graph is drawn. Each vertex in the graph holds the

	Words			Score
	<i>Society</i>	<i>Civilization</i>	<i>Culture</i>	
I n t e r p r e t a t i o n s	s1	s5	s6	2
	s1	s5	s7	0
	s1	s5	s8	0
	s1	s5	s9	0
	s1	s5	s10	0
	s2	s5	s6	2
	s2	s5	s7	0
	s2	s5	s8	0
	s2	s5	s9	0
	s2	s5	s10	0
	s3	s5	s6	2
	s3	s5	s7	0
	s3	s5	s8	0
	s3	s5	s9	0
	s3	s5	s10	0
	s4	s5	s6	4
	s4	s5	s7	2
	s4	s5	s8	2
	s4	s5	s9	2
	s4	s5	s10	2

Figure 38: Of all the different interpretations of the sequence, one is singled out in this example, marked by the bold frame. The boldfaced and enlarged senses indicate the existing relationships between the meanings. The interpretation that scores the highest is the one that involves meanings s4, s5 and s6, since all three are related.

sense_ID (or rather the shorthand version for it that I devised) associated with the meaning. Also in each vertex are the words that have this meaning in this interpretation. So if two words had had the meaning s6, they would both have been presented in the s6 vertex. The edges represent lexical relations and are labelled “h”, ”m” or “a” for hyponymous, meronymous and antonymous respectively.

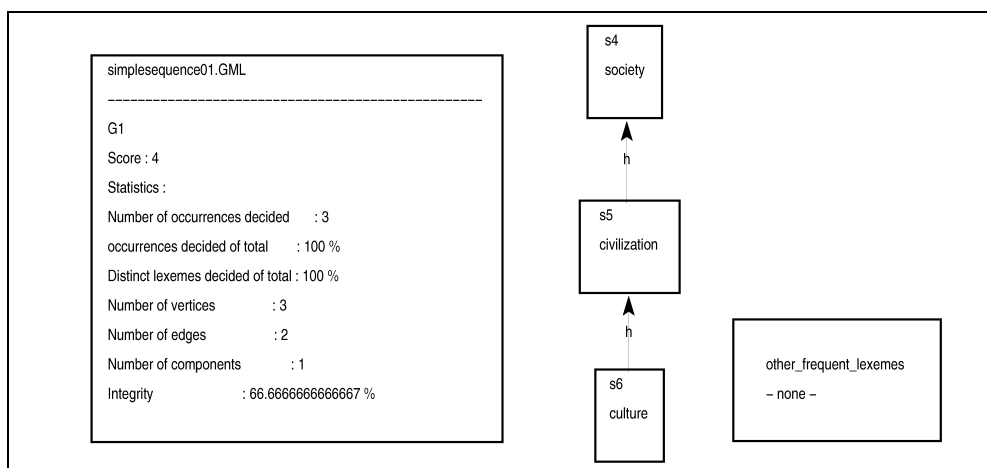


Figure 39: The graphical representation of the simple example sequence “*society civilization culture*”.

While the graph itself remains the object of importance I made the system output a couple of extra vertices in addition to the graph itself.

First I included a meta-vertex meant to hold un-interpreted words that occur frequently in the sequence. This is to say that even if the system were not able to assign a meaning to a particular word the mere fact that it occurs ten times in the sequence indicates that it just might have some significance to the semantic context of the original text. It might be a proper name or even a new concept being coined that doesn’t appear in the MRD. This vertex does not, however, affect the scoring in any way, but was included for its potential importance in future work. In this graph all instances in the sequence are interpreted so the “other_frequent_lexemes” meta-vertex is left empty.

Secondly, I have meta-vertex holding information about the data-file, the score of this particular graph and also various statistical information that I found useful for evaluating the scoring-system. An explanation of this information seems in order.

I start by making a reference to the GML file holding the graph in question. I name GML files according to the original data files holding the sequence being skimmed. I number alternative interpretations in successive order.

Then follows the score assigned by the scoring system to the graph, here the score is 4.

Finally the statistical information includes:

- A count of individual instances in the sequence that were successfully assigned a meaning by this interpretation.
- An indication of the percentage of such instances out of the total number of instances in the sequence.
- A similar percentage for distinct lexemes. So if a sequence held four instances, say “*book dictionary Oxford dictionary*”, three of which were successfully assigned a meaning *book, dictionary, dictionary* while the last *Oxford* is undecided in this interpretation, the system would have interpreted three occurrences. That is 75%

of four occurrences in all, while it would have interpreted only two lexemes making for 67% of three lexemes in total.

- The number of vertices and edges relate to the number of senses and their interrelationships as already described.
- The number of connected components count the number of distinct connected sub-graphs that are present in the graph - the rationale being that the fewer components in the graph the more focused the semantic context of the sequence. So if a graph holds relatively few components this graph should also score relatively high.
- Finally, I present a measure of how tight the focus is. Integrity is a percentage ratio of edges relative to the maximally possible for an undirected graph with that many vertices. Here we have two edges, whereas the maximally possible number of edges in an undirected graph of three vertices is three - making for 67% integrity. 100% integrity occurs very rarely. The integrity measure was included as an evaluation tools for the scoring system particularly to see if there is a connection between subjective preference of interpretation and degree of “connectedness” of the respective graphs produced by the skimmer.

7.2.2 Other output files

In addition to the GML files presenting each partial interpretation as I just described the prototype system outputs three more files for each input sequence. I shall describe them briefly here.

The first additional file I include is a meta-file (i.e.: a file describing another file but yet distinct from it) in the GML format for each input sequence.

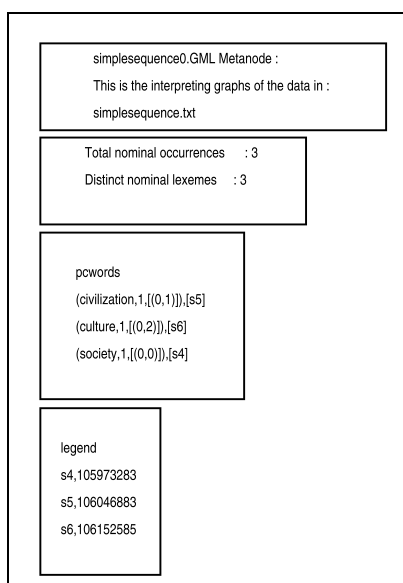


Figure 40: One meta-file GML for each sequence

This GML-file is meant to accompany each series of interpreting graphs but put in its own file in order to avoid cluttering up the graphs. It

sums up the number of instances as well as the number of distinct lexemes in the sequence. The File has four vertices, the first two of which I trust are self-explanatory. The vertices “**pcwords**” and “**legend**” need explanation.

Pcwords – possibly connected words: This vertex lists those nouns along with the number or times it occurs in the sequence, their positions and those of their meanings that are connected to each other. I.e.: the word “culture” occurs as the third instance in the first paragraph of the sequence. It has position (0,2) I.e.: paragraph 0 position 2. Of its meanings only s6 has relations to meanings of the other words in the sequence. Only words that have at least one meaning relating to meanings of other words in the sequence appear in “**pcwords**”.

Legend: This vertex holds the mapping between the relevant WordNet Synset ID’s and my shorthand version. This mapping is generated a new for each sequence fed to the system. One particular sequence will get the same mapping on each such occasion.

```
Glossary for datafile : simplesequence.txt

glossary
s1, (the state of being with someone; "he missed their company"; "he enjoyed
the society of his friends")
s10, (the raising of plants or animals: "the culture of oysters")
s2, (the fashionable elite)
s3, (a formal association of people with similar interests; "he joined a golf
club"; "they formed a small lunch society"; "men from the fraternal order
will staff the soup kitchen today")
s4, (an extended social group having a distinctive cultural and economic
organization) (continued..)
s5, (a society in an advanced state of development)
s6, (a particular civilization at a particular stage)
s7, (all the knowledge and values shared by a society)
s8, (the tastes in art and manners that are favoured by a social group)
s9, ((biology) the growing of micro organisms in a nutrient medium (such as
gelatin or agar); "the culture of cells in a Petri dish")
```

Figure 41: The WordNet gloss-description for the different senses of the words in the sequence. I make the sense references in boldfaced types to help visual distinction between the entries.

The Second additional file I include is a Gloss file (GLS), relating a short textual description of the senses that *were in consideration* for the words of the sequence. It is fairly easy to realise just what is hidden behind the Sense_ID’s when we get a hint. The Gloss is taken from the GLOSS database of WordNet and provided here as a service to non-machine readers.

Finally I include the systems non-graphical output file (OUT). This was actually the only output of the system before I decided to use the Graphlet application. As such it can be seen as an alternative to the graphical representation, that is - an alternative to actually drawing the graphs. It should be fairly self-explanatory by now.

Note, however, that alternative interpretations would all be present in this one file. I.e.: had there been an alternative partial interpretation for this

sample sequence there would have been a G2 paragraph following the G1 paragraph - and so on⁹.

Also this file holds a *pcenses* structure. This structure is really nothing else but the transposition of the *pcwords* already mentioned.

Lastly there is mention of a structure named *pedges*. This contains the relationships involving the alternative senses or the words and is what I referred to as the set R_0 in the analysis of chapter 6. These are edges that candidate for inclusion in an interpreting graph and hence the name *possible edges*. Each edge is represented by the kind of relation, i.e.: “h”, ”m” or “a”, and the pair of vertices that are involved. Furthermore each edge is associated with its respective score.

When describing and discussing the various examples and experiments I will for the most part refrain from displaying the OUT-file because the information contained here is represented in full by the other files. Still I illustrate it for the sake of completeness in Figure 42 below.

⁹ Of course, there are alternative interpretations. - Nineteen of them in fact, as I have already pointed out. In order for alternative graphs to be reported, however, I require that they either decide at least the same number of lexemes or score at least the same. Examples offering several alternative interpretations of the same sequence will be shown and discussed shortly. Finally, all of those alternative interpretations that involve at least one lexical relationship are in fact present in the internal structures of the system.

```

simplesequence0.out

pcsenses
s4,1,[(society,1,[(0,0))]]
s5,1,[(civilization,1,[(0,1))]]
s6,1,[(culture,1,[(0,2))]]

pcwords
(civilization,1,[(0,1)],[s5]
(culture,1,[(0,2)],[s6]
(society,1,[(0,0)],[s4]

legend
s4,105973283
s5,106046883
s6,106152585

pedges
e(2,h,(s5,s4))
e(2,h,(s6,s5))

Total nominal occurrences in data : 3
Total distinct nominal lexemes in data : 3
-----
G1
Score : 4

relationships
e(2,h,(s5,s4))
e(2,h,(s6,s5))

interpretations
s4,[society]
s5,[civilization]
s6,[culture]

other_frequent_lexemes
- none -

Statistics :
Number of occurrences decided : 3
Occurrences decided of total : 100 %
Distinct lexemes decided of total : 100 %

Number of senses : 3
Number of relationships : 2
Number of connected components : 1
Integrity : 66.6666666666667 %
-----

```

Figure 42: The alternative output format with no graphical graphs.

7.2.3 Twisting it a bit

So this was how the prototype works on a very simple and strictly controlled example sequence with no contradictions. In order to further demonstrate how the prototype system is working let's add a little spice to the example before we go to the corpus sequences of CIVIII. Spice in this respect is contradictory interpretations that compete for highest score. One of the polysemous words in the simple sequence was *society*, so let's experiment with a sequence that must consider other interpretations of *society* than the one that scored the best before. The following sequence should promote the ESOTERIC_SOCIETY meaning of *society*.

```
p.
association.
society.
chapter.
```

Figure 43: Another sequence involving the word *society*, Will this make for another reading of the word ?

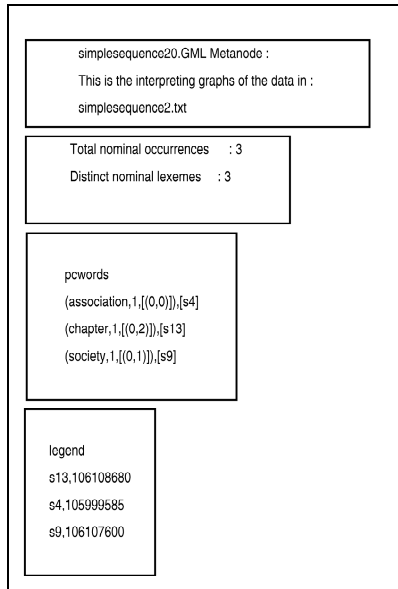


Figure 44: The meta-file of simplesequence2.txt

In the *pcwords* vertex we see that each word in the sequence has one meaning that is connected to a meaning of another word in the sequence. Furthermore we see that no word has more than one such meaning. This makes for one partial interpretation that happens to be complete as well, and with no competing alternatives. The system confirms this by reporting the following graph.

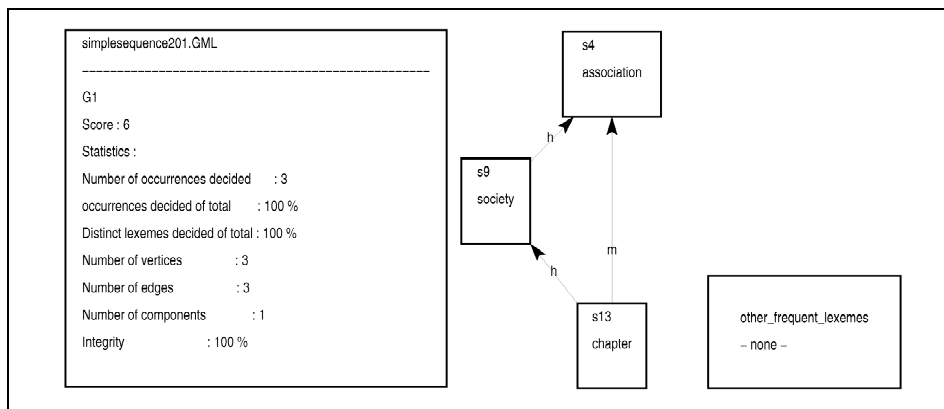


Figure 45: The interpreting graph of "simplesequence2". Recall that "m" refers to a meronymous relationship between meanings indicating *part-of*, *member-of* or *portion-of* relations. "h" still refers to hypernymous relationships indicating *is-a* or *is-a-kind-of* relations.

It should be apparent that in this interpretation of the words occurring in the sequence a CHAPTER is part of an ASSOCIATION and is a kind of SOCIETY as well. In turn SOCIETY is a kind of ASSOCIATION. Each word appears only once in the sequence and so adds 1 to the score of their respective lexical vertices as we could call them. Each vertex is the ending point of two relational edges and so all score twice. Hence the total score of 6 for this interpretation. Also this graph is an example of the relatively rare interpreting graphs with 100% integrity. To get an idea of what concepts hide behind the sense-references lets take a look at the gloss file for the sequence.

```
Glossary for datafile : simplesequence2.txt

glossary
s1, (the state of being connected together as in memory or imagination; "his
association of his father with being beaten was too strong to break")
s10, (an extended social group having a distinctive cultural and economic
organization)
s11, (a distinct period in history or in a person's life; "the industrial
revolution opened a new chapter in British history"; "the divorce was an
ugly chapter in their relationship")
s12, (an ecclesiastical assembly of the monks in a monastery or even of the
canons of a church)
s13, (a local branch of some fraternity or association; "he joined the
Atlanta chapter")
s14, (a series of related events forming an episode; "a chapter of
disasters")
s15, (a subdivision of a written work; usually numbered and titled; "he read
a chapter every night before falling asleep")
s2, (a social or business relationship: "a valuable financial affiliation";
"he was sorry he had to sever his ties with other members of the team";
"many close associations with England")
s3, (any process of combination (in solution) that depend on relatively weak
chemical bonding)
s4, (a formal organization of people; "he joined the Modern Language
Association")
s5, (the process of bringing ideas or events together in memory or
imagination; "conditioning is a form of learning by association")
s6, (the act of consorting with or joining with others; "you cannot be
convicted of criminal guilt by association")
s7, (the state of being with someone; "he missed their company"; "he enjoyed
the society of his friends")
s8, (the fashionable elite)
s9, (a formal association of people with similar interests; "he joined a golf
club"; "they formed a small lunch society"; "men from the fraternal order
will staff the soup kitchen today")
```

Figure 46: The GLS file for "simplesequence2".

Recall that the mapping between Synset_Id of the MRD and the shorthand version for it is generated a new. Thus the meaning that was assigned to *society* in the previous example is named s10 in this example instead of s4. However, we see that *society* is assigned a totally different meaning in this example, namely s9 that refers to a much closer relationship between its members. Also the explanatory wording for the meanings assigned to *association*, and *chapter*, (s4 and s13 respectively) conforms nicely to a rather more esoteric organisation than before.

7.2.4 Contradicting interpretations

I just presented a second sequence describing a quite different context of the word *society*. What will be the interpretational consequences if we

join the two sequences into one? The new sequence is presented in Figure 47.

```
p.
civilization.
culture.
society.
association.
chapter.
```

Figure 47: “simplesequence4.txt”. This sequence contains contradictory contexts for the word *society*

Note, that still each word is present only once in the sequence. The system quite obviously has a hard time to decide what to make of the word *society* and this is reflected in its response, depicted in Figure 48, Figure 49 and Figure 50.

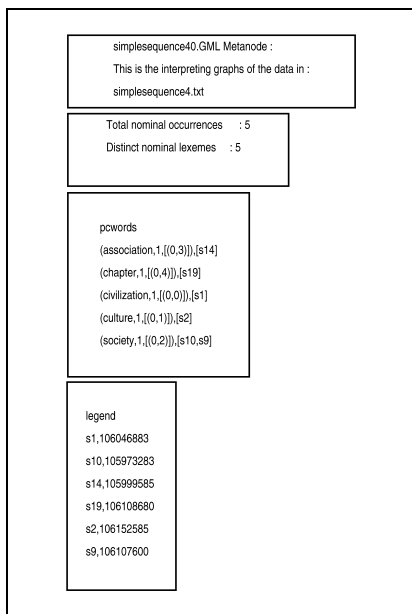


Figure 48: The GML metafile for the new sequence. Observe how the ambiguity of *society* in this example is spelled out in the *pcwords* vertex, listing two alternative connected meanings for the word – both s10 and s9.

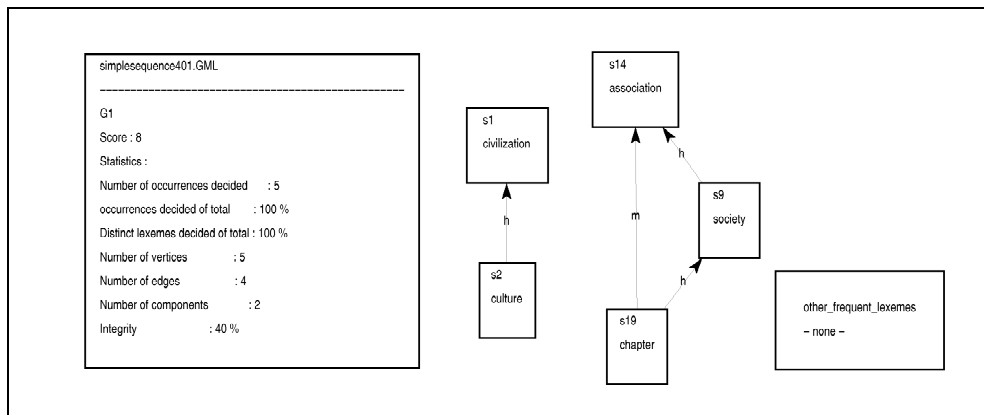


Figure 49: The first interpreting graph of “simplesequence4.txt” places *society* in the organisation camp so to speak. This interpretation scores 8, which we can confirm by counting edges and adding 2 points for each.

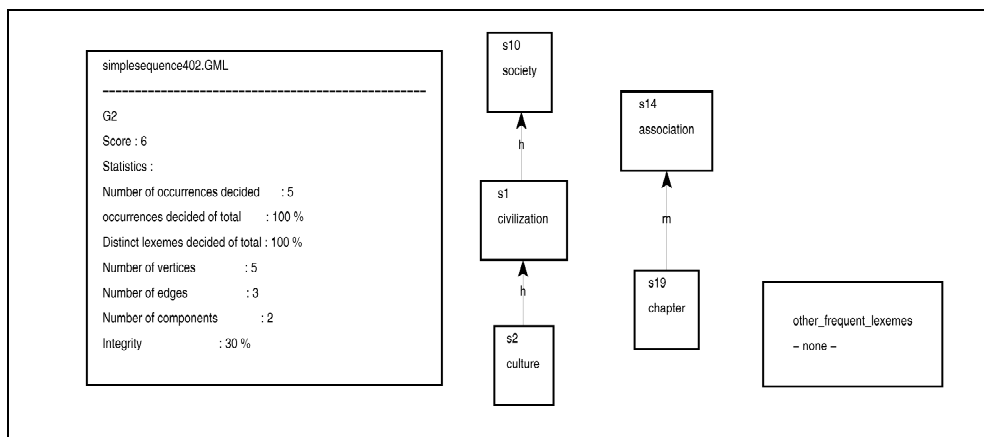


Figure 50: The second interpretation of ”simplesequence4.txt” retains *society* among the demographic concepts. Counting edges this interpretation scores only 6.

Glossary for datafile : simplesequence4.txt

glossary

s1, (a society in an advanced state of development)

s10, (an extended social group having a distinctive cultural and economic organization)

s11, (the state of being connected together as in memory or imagination; "his association of his father with being beaten was too strong to break")

s12, (a social or business relationship: "a valuable financial affiliation"; "he was sorry he had to sever his ties with other members of the team"; "many close associations with England")

s13, (any process of combination (in solution) that depend on relatively weak chemical bonding)

s14, (a formal organization of people; "he joined the Modern Language Association")

s15, (the process of bringing ideas or events together in memory or imagination; "conditioning is a form of learning by association")

s16, (the act of consorting with or joining with others; "you cannot be convicted of criminal guilt by association")

s17, (a distinct period in history or in a person's life; "the industrial revolution opened a new chapter in British history"; "the divorce was an ugly chapter in their relationship")

s18, (an ecclesiastical assembly of the monks in a monastery or even of the canons of a church)

s19, (a local branch of some fraternity or association; "he joined the Atlanta chapter")

(continued..)

<p><u>s2</u>, (a particular civilization at a particular stage) <u>s20</u>, (a series of related events forming an episode; "a chapter of disasters") s21, (a subdivision of a written work; usually numbered and titled; "he read a chapter every night before falling asleep") s3, (all the knowledge and values shared by a society) s4, (the tastes in art and manners that are favored by a social group) s5, ((biology) the growing of microorganisms in a nutrient medium (such as gelatin or agar); "the culture of cells in a Petri dish") s6, (the raising of plants or animals: "the culture of oysters") s7, (the state of being with someone; "he missed their company"; "he enjoyed the society of his friends") s8, (the fashionable elite) <u>s9</u>, (a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today")</p>
--

Figure 51: the glossary for "simplesequence4.txt". The senses included in the two interpretations have been underlined. Senses s9 and s10 are the competing meanings of the word *society*.

As can be seen, there is an envious struggle between two components for the inclusion of *society*. In fact this serves as a good illustration of just what is represented by the graphs and their components. I propose that any text producing this sequence of nouns must in fact involve two separate contexts - hence the two components.

One context relates to demographics that deal with the large-scale organisation of the people of the world.

Another context relates to the more intimate organisation of individuals having personal goals in common.

Of course the two context shares the fact that they both deal with the organisation of people. The difference lies in the level of abstraction of the respective ways of organising. The problem with this sequence is the use of the word *society* that has meaning in both the contexts, but a different one in either. Referring back to *The Cooperative Principle* of Grice, this would make the author of the "original text" guilty of breaking the maxim of manner that says that one should "*be perspicuous*" and in particular that one should "*avoid ambiguity*". I am a little more lenient than Grice in that words are of course allowed to have different meanings just not, as it is, within the same paragraph.

Returning to the two interpreting graphs G1 and G2, we see that even though both offer complete interpretations of the sequence in question, they do not score the same. The fact that they both offer disambiguation of the same number of words is the reason why they are both reported, instead of just the high scoring one. Because the scoring is based on the inclusion of edges, the meronymous edge in G1 of Figure 49 causes this interpretation to outscore G2 of Figure 50.

It is a valid argument that the two interpretations should rightly score the same since they are both complete interpretations of the sequence. It is, of course, hard to decide what the proper interpretation is when the sequence is an artificial one like this - constructed for the purpose of this example. There is no original text that generated this particular sequence. I am sure a couple of examples can be construed over the sequence, that each promotes one interpretation over the other.

It remains my argument though, that the more interrelationships between the concepts of the interpretation - the more focused the context(s). Once again referring to **Grice**, (Grice, H. P. 1975) The *maxim of manner* dictates that, in order to make sense of the mess, I should go with the more focused of the two. Thus I for now maintain that the scoring scheme, as it is, works as it should.

7.3 Experimenting with corpus data

Having described the basic function of the experimental prototype through carefully constructed and firmly controlled sequences I will now turn to data as it could occur in actual natural language text.

The Civilization III manual offers plenty of such informative textual data. In fact the Civilization III manual represents one of the most thorough and detailed computer game manuals that I have come across. The printed version of the manual involves a paperback book of over 200 pages. For the first experiments with my prototype *skimming*-program I have chosen the introductory Chapter 2 of the manual. This chapter introduces a wide variety of concepts and situations to help the reader appreciate all the nuances of the game itself. Because of the introductory nature of this chapter, many different and perhaps contrasting contexts are abundant and so the chapter should make for some interesting experiments.

As for the purpose of the experiments, I will focus on the following two main interests: I. how well does the prototype succeed in establishing recognizable contexts for the textual paragraphs of the Corpus? and II., how do the contexts of different paragraphs relate to each other? Each of these two I will divide into the following sub questions:

- I. Study the graphs that are generated by the prototype :
 - a) Do they properly represent their respective textual paragraphs ?
 - b) There will be examples of erroneous interpretations, i.e.: established lexemes that do not subjectively belong in the context of the textual paragraph :
 - i. How many such errors occur relative to the number of interpretations in the graph?
 - ii. What is the reason for these errors – the algorithm/scoring-system, the design of WordNet or are they simply coincidental or a result of specific textual style ?
- II. Study the relationships between the textual paragraphs and the assignment of meanings to words, specifically across paragraph boundaries. Certainly the author of the manual organised the paragraphs in order to best reflect what he/she wanted to tell the reader. In this respect it is interesting to experiment with paragraph boundaries. The prototype does not employ means to automatically compare contexts of separate sequences as it is. It is however possible to experiment with the possible outcomes of

such reasoning means. Representing separate paragraphs as separate sequences, first apply the skimmer to one sequence then to the other sequence and finally to the sequence resulting from concatenating the two. The concatenation of sequences will correspond to disregard of the original boundaries separating the paragraphs and regard them as one paragraph. This obviously will produce three sets of interpreting graphs. Under the assumption that context boundaries reflect paragraph boundaries, comparing the three interpretations it may be possible find signals that indicate whether the two portions are better interpreted in isolation or as describing the same context. Consider for example the following points.

- a) Lexemes that prevail in the conjunction of two paragraphs might indicate that the conjunction is a *continuation* of the two paragraphs, i.e.: that they maybe have the same context.
- b) Lexemes that are established in such a conjunction, but were not established in the either of the respective paragraphs in isolation, might indicate that the conjunction is an *expansion* of the two participating contexts, i.e. both participating contexts prevail and new concepts are identified as well.
- c) Lexemes present in either participating context in isolation but not in the conjunction of the two might indicate contradicting contexts and that the two contexts, and respective paragraphs, should be regarded as distinct (*distinction*).
- d) A word having different meanings in the interpretations of two paragraphs but is only one when joining the two paragraphs into one might indicate *distinction* or that the conjunction properly decides the meaning of the word (*refinement*).
- e) Lexemes of a given paragraph that are replaced when joining that paragraph with another might indicate *distinction* or *refinement*.

In order to address these questions I will conduct the following experiments with the skimming prototype applied to the CIVIII corpus:

- 1) Treat each paragraph of the corpus (Figure 52) in isolation and provide answers to questions in part I of the experimental interests above as they present themselves.
- 2) Regard the typographical structure of the text in Figure 52. It looks like the author intended that the paragraph be interpreted as 0(1,2,3,4,5),6,7,8. In order to see if a similar structure is reflected in the interpretations of the prototype, join the following pairs of paragraphs (0,1), (0,2), (0,3), (0,4), (0,5), (0,6), (0,7), (0,8) and provide answers to the question in part II of the experimental

interests above as they present themselves. Does the pairs of the first five conjunctions relate closer to each other than the pairs of the last three ?

- 3) Join the paragraphs 0,1,2,3,4,5 and compare to results of experiment 1. Are there any indications as to whether they should be interpreted in conjunction or in isolation ?
- 4) Join all nine paragraphs into one and compare to results of the earlier experiments. Is there any indication as to the structure of the contexts ?

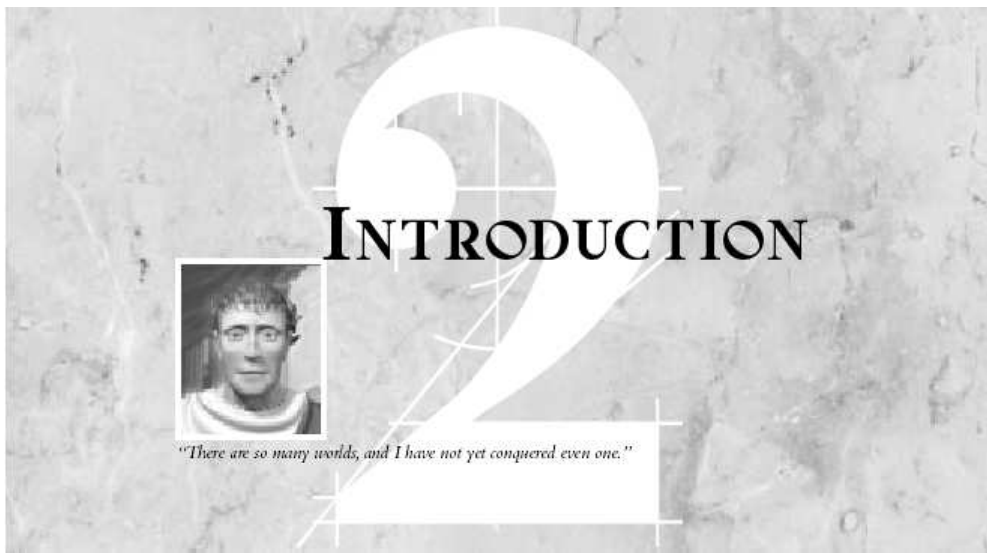
Apart from adhering to the presented guidelines I must admit that I do not have a strictly objective and consequent method of evaluation at my disposal. The results of the experiments will be discussed as they appear to me and compared to what I will refer to as a *subjective* result. The *subjective interpretation* of a given sequence is intended to refer to that interpretation that an average human reader would most likely produce for the sequence. In other words, I compare the interpretations made by the algorithm to the interpretation that I think is the proper or most plausible interpretation for the given sequence. This is of course obviously in need of further experiments involving test persons that are considerably more removed from the project than myself. Such full scale blind tests is unfortunately not going to be carried out in the context of this project and I will have to make do with my own, hopefully sound judgement. Still, the project being of an prototypical experimental nature, I am confident that the proposed “method” will at least provide a reasonably detailed picture of what problems need to be addressed in future work and how the skimmer perform on a small scale.

Each of these four experiments will generate their own “legend” of the mapping between shorthand notation and the actual Synset_ID of the MRD (see also “LEGEND” in section 7.2.2). In order to avoid confusion of reference I will for the experiments 1,2 and 3 refer to senses using the mapping from experiment 4 in addition to the mapping of the respective experiments themselves. To distinguish between mappings I will refer to them as follows :

	Exp 1 Ref.	Exp 4 ref.	
<i>objective</i>	s23	s14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"

In the figure above (Exp 1 Ref) refers to the shorthand Synset_ID that was assigned to sense of objective in experiment 1, while (Exp 4 Ref) refers to the shorthand Synset_ID that sense received in Experiment 4. So s23 and s14 above both refers to the same sense just in different applicative runs of the prototype. The references in the graphs themselves refer to the shorthand Synset_ID of the experiment itself, but Exp 4 references makes for the easiest comparison of the word senses from one experiment to another.

The graphs for the all the experiments can be found in their respective subsections of the appendix A-1 through A-4. In the appendices I only included the glossary file for experiment 4 since it fully covers all four experiments. Including the glossary files for experiments 1, 2 and 3 would involve significant redundancy and loss of clarity.

	
0	<p>Five Impulses of Civilization</p> <p>There is no single driving force behind the urge toward civilization, no one goal toward which every culture strives. There is, instead, a web of forces and objectives that impel and beckon, shaping cultures as they grow. In the <i>Civilization III</i> game, five basic impulses are of the greatest importance to the health and flexibility of your fledgling society.</p>
1	<p>Exploration</p> <p>An early focus in the game is exploration. You begin the game knowing almost nothing about your surroundings. Most of the map is dark. Your units move into this darkness of unexplored territory and discover new terrain; mountains, rivers, grasslands, and forests are just some of the features they might find. The areas they explore might be occupied by minor tribes or another culture's units. In either case, a chance meeting might provoke a variety of encounters.</p>
2	<p>Economics</p> <p>As your civilization expands, you'll need to manage the growing complexity of its production and resource requirements. Adjusting the tax rates and choosing the most productive terrain for your purposes, you can control the speeds at which your population grows larger and your cities produce goods. By setting taxes higher and science lower, you can tilt your economy into a cash cow. You can also adjust the happiness of your population. Perhaps you'll assign more of your population to entertainment, or you might clamp down on unrest with a larger military presence. You can establish trade with other powers to bring in luxuries and strategic resources to satisfy the demands of your empire.</p>
3	<p>Knowledge</p> <p>On the flip side of your economics management is your commitment to scholarship. By setting taxes lower and science higher, you can increase the frequency with which your population discovers new technologies. With each new advance, further paths of learning open up and new units and city improvements become available for manufacture. Some technological discoveries let your cities build unique Wonders of the World.</p>
4	<p>Conquest</p> <p>Perhaps your taste runs to military persuasion. The <i>Civilization III</i> game allows you to pursue a range of postures, from pure defense through imperialistic aggression to cooperative alliance. One way to win the game is to be the last civilization standing when</p>

	the dust clears. Of course, first you must overcome both fierce barbarian attacks and swift sorties by your opponents.
5	<p>Culture</p> <p>When a civilization becomes stable and prosperous enough, it can afford to explore the Arts. Though cultural achievements often have little practical value, they are frequently the measure by which history—and other cultures—judge a people. A strong culture also helps to build a cohesive society that can resist assimilation by an occupying force. The effort you spend on building an enduring cultural identity might seem like a luxury, but without it, you forfeit any chance at a greatness other civilizations will respect.</p>
6	<p>The Big Picture</p> <p>A winning strategy is one that combines all of these aspects into a flexible whole. Your first mission is to survive; your second is to thrive. It is not true that the largest civilization is necessarily the winner, nor that the wealthiest always has the upper hand. In fact, a balance of knowledge, cash, military might, cultural achievement, and diplomatic ties allows you to respond to any crisis that occurs, whether it is a barbarian invasion, an aggressive rival, or an upsurge of internal unrest.</p>
7	<p>Winning</p> <p>There are now more ways of winning the game. You can still win the Space Race with fast research and a factory base devoted to producing spacecraft components. You can still conquer the world by focusing on a strong military strategy. If you dominate the great majority of the globe, your rival may well give in to your awesome might. In addition, there's a purely Diplomatic means of success; if you're universally renowned as a trustworthy peacemaker, you can become head of the United Nations. Then there's the challenge of overwhelming the world with your Cultural achievements—not an easy task. Finally, of course, is perhaps the most satisfying victory of all—beating your own highest Historic Civilization Score or those of your friends. See Chapter 14: Winning the Game for an in-depth analysis of the scoring system.</p>
8	<p>The Documentation</p> <p>The folks who make computer games know that most players never read the manual. Until a problem rears its head, the average person just bulls through by trial and error; it's part of the fun. When a problem does come up, this type of player wants to spend as little time in the book as possible, then get back to the game. For those of you who are looking for a quick fix, Chapter 15: Reference: Screen by Screen is the place to go. For the rest of you, we've tried to organize the chapters in the order that you'll need them if you've never played a <i>Civilization</i> game before. If you're new to the game, the sidebars on concepts should help you understand the fundamentals of the game. The Readme file on the CD-ROM has the rundown on the very latest changes, things that didn't make it into this manual. (Due to printing and binding time, the manual has to be completed before final tweaks are made.)</p> <p>Last but not least, the <i>Civilization III</i> game continues the tradition of including a vast compendium of onscreen help. Click on the Civilopedia icon (the book near your advisors) or on any hyperlinked text in the game to open the Civilopedia. This handy reference includes entries describing all the units, improvements, governments, terrain, general game concepts, and more—everything you could want to know about the <i>Civilization</i> world. The entries are hyperlinked so you can jump from one to another with ease.</p>

Figure 52: The first nine paragraphs of Chapter 2 "Introduction" of the Civilization III manual. This is the textual data for the starting experiments with the prototype program. For ease of reference paragraph numbers has been noted in the left column of the table. The chapter goes on for several paragraphs of a subsection devoted to interface conventions but there is plenty of data in these nine paragraphs for carrying out my starting experiments.

7.4 Experiment 1 - isolated paragraphs

The first simple experiment is to apply the skimming prototype to each of the nine paragraphs in isolation. The prototype should provide a readily identifiable context for each paragraph and also the “proper” reading for some of the words of the particular paragraph.

Note, that in the discussion of the experimental results I employ two slightly different uses of the term interpretation. One that serves as a shorthand equivalent of the term “interpreting graph” and one that refers to the pairing of a particular word to a particular meaning. I realize that this in essence violates the maxim of clarity, but I trust that the context will provide the needed clarity.

7.4.1 Paragraph 0

The first paragraph contains three distinct components assigning meanings to seven of the fourteen nouns. the words as follows :

	Exp 1 Ref.	Exp 4 ref.	
<i>civilization</i>	s30	s693	a society in an advanced state of development
<i>culture</i>	s25	s292	a particular civilization at a particular stage
<i>force</i>	s20	s312	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
<i>goal</i>	s4	s4	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
<i>impulse</i>	s42	s19	an impelling force or strength; "the car's momentum carried it off the road"
<i>objective</i>	s23	s14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>society</i>	s55	s300	an extended social group having a distinctive cultural and economic organization

Figure 53 Interpretations of the words in paragraph 0 and their glossary text.

Reading the paragraph (see Figure 52), the interpretations of *goal* and *objective* seems very reasonable to me. Also the interpretations of *civilization*, *culture* and *society* seem very close to what must be the intended meaning of the paragraph.

The words *force* and *impulse* has been boldfaced in the table above. This indicates that I think that these particular interpretations offered by the system are questionable. Concretely, the words have been interpreted with reference to some PHYSICAL_FORCE. I think the FORCE at play in the paragraph is a rather more psychological one.

Among the words of the paragraph that do not get interpreted, especially the word *urge* stands out as having a meaning with obvious semantic relations to the kind of FORCE and IMPULSE that I think should have been present in the interpretation. *Urge* is not among the Pwords of the graph however, indicating that no direct relation exists between any two senses of those three words in the MRD. This may be a consequence of the fine-grained-ness of WordNet. More on that in the next chapter.

Summing up, all of the Pwords of the graph appear in the interpretation, and this indicates that indeed the paragraph does not contain

contradictive lexemes. The components of the graph to me do represent a reasonable summary of what the paragraph is about. All in all the scoring system does well apart from the force-impulse component that looks like it does not belong in this particular context.

The actual files for this experiment can be found in figure A-1.1 of the appendix A-1. See appendix A-4 for glossary.

7.4.2 Paragraph 1

The interpretations offered of the second paragraph are lacking a lot of information. The paragraph holds reference to many different kinds of terrain, it refers to exploration, maps and the like. I would have liked a clear indication of a context flavoured by geographical concepts. Instead, we see a transferred meaning of *area* and *territory* referring to

AREA_OF_KNOWLEDGE_AND_STUDY.

The absence of words like *mountain*, *river*, *grassland* and *forest* in relation to *terrain* and *territory* I deem to be a consequence of intervening relationships between these concepts in the MRD. There may be, for instance, a link is present in the MRD that connects a meaning of *terrain* to a notion like GEOGRAPHICAL_TERRAIN and from this to the various kinds of such terrain types.

The very fine-grained distinctions of WordNet should be a positive but in this case - and others like it - it prevents the algorithm from making the right choices and instead leads to wrong answers.

I am almost certain that, if the prototype was to be expanded to take into regard also relationships with intervening links, instead of only direct links, this problem would be eliminated. Such a modification of the algorithm is likely to grossly worsen its complexity, due to a greatly increased number of edges in the chunk. (Trying this out represents an interesting and necessary experiment for future work).

	Exp 1 Ref.	Exp 4 ref.	
<i>area</i>	s93	s57	a subject of study; "it was his area of specialization"; "areas of interest include..."
<i>territory</i>	s79	s44	an area of knowledge or interest; "his questions covered a lot of territory"

Figure 54 attempted interpretations of the interrelated lexemes of paragraph 1.

Summing up, the skimming algorithm fails to offer any felicitous interpretation of the words in paragraph 1. Also the resulting context is decidedly misleading.

The full set of files and tables of this paragraph can be found in figure A-1.2 of the appendix A-1. See appendix A-4 for the glossary.

7.4.3 Paragraph 2

There are three alternative interpretations of this paragraph. Since, however, the first one scores better than the two latter, this is the one that should provide the better interpretation. Let us see how it turns out.

Paragraph 2 deals with various ways to manipulate the economy of the in game society.

In this respect *science* and *taxes* are reasonably well interpreted. In the paragraph, *speed*, relates to *the growth rate of the population* and thus the s114 interpretation of *speed* seems felicitous as well and in direct relation with the s108 sense of *rate*. Unfortunately *rate* in the paragraph is the TAX_RATE ,i.e. a meaning much closer to that of s105. To spoil the milk even more the close relation between the s107 sense of *rate* and the *tax* lexeme causes it to get chosen for the best interpretation, since *tax* is mentioned several times in the paragraph.

Finally the word *power* in the paragraph occurs more or less as a synonym to *culture*, *civilization* or *government*. A sense of *power* that does not appear in this interpretation at all. Instead *power*” gets a transferred reading relating to THE_MENTAL_ABILITIES_OF_INDIVIDUALS.

Summing up, the skimming procedure offers reasonable good interpretations for *science* *speed* and *tax*, while the interpretations offered for *power* and *rate* are not as good. The resulting context does not quite cover the paragraph. It is on the other hand not easy to get hold of just what goes on in there from the noun-list alone even for a human. But at least some hint at an economic/financial context is present. Perhaps there is indication here, that the scoring system need fine-tuning or we have a series of unfortunate coincidences in the rather short paragraph. More on this in the next chapter.

	Exp 1 Ref.	Exp 4 ref.	
<i>power</i>	s211	s145	possession of the qualities (especially mental qualities) required to do something or get something done; "danger heightened his powers of discrimination"
<i>rate</i>	s156	s107	(British) a local tax on property (usually used in the plural)
	s157	s108	the relative speed of progress or change; "he lived at a fast pace"; "he works at a great rate"; "the pace of events accelerated"
	s154	s105	a magnitude or frequency relative to a time unit; "they traveled at a rate of 55 miles per hour"; "the rate of change was faster than expected"
<i>science</i>	s174	s181	ability to produce solutions in some problem domain; "the skill of a well-trained boxer"; "the science of pugilism"
<i>speed</i>	s164	s114	a rate (usually rapid) at which something happens; "the project advanced with gratifying speed"
	s162	s112	distance travelled per unit time
<i>tax</i>	s171	S178	charge against a citizen's person or property or activity for the support of government

Figure 55 Interpretations of the words in paragraph 2 of the corpus text.

The interpreting graphs and tables relating to this paragraph can be found in A-1.3 of Appendix A-1. See appendix A-4 for the glossary.

7.4.4 Paragraph 3

The skimming prototype fails to find any relationship between the words in the paragraph headlined KNOWLEDGE in Figure 52. There are

words in this paragraph that to me obviously relate to the pursuit of knowledge however (- *knowledge, scholarship, science, technology, discovery*). The fact that the system cannot realize these relationships must rely on intervening concepts in the MRD, as mentioned before.

The ability to recognise a relationship between two concepts via one, that is not itself realised explicitly in the text (I will call such a concept **implicit**), really would help here, I think.

Since the prototype is presently restricted to take into regard only direct relationships between concepts that are realised in the text, the interpreting graph of paragraph 3 is however empty.

Appendix A-1.4 has the particulars of this paragraph. See appendix A-4 for the glossary.

7.4.5 Paragraph 4

This paragraph introduces the military aspects of the game, referring to *conquest, defense, aggression, alliance, attack, sortie* and *opponent*. Again, however, the graph is empty: the relationships between these concepts are not recognised by the skimming algorithm – relationships that are immediately apparent to a human reader. In particular I would think that *attack* and *defense* would be immediately related by antonymy and it is a bit surprising that the algorithm does not realize this. Maybe it relies on the American/English spelling of defense, i.e. with an “s”, but then again the MRD, WordNet, was developed in the States.

See Appendix A-1.5 for details of this paragraph. See appendix A-4 for the glossary.

7.4.6 Paragraph 5

This paragraph reintroduces the many cultural aspects of the Civilization III game. The skimmer produces one interpreting graph, scoring 11. The graph beautifully continues the cultural component of paragraph 0 and adds to it the *achievement* and *arts* components. The main reason this goes so well is of course that the author chose to refer to the core concepts of *society, civilization* and *culture* recognised in paragraph 0. This is in perfect accord with the maxims of **Grice** and obviously eases the task of orientation across paragraph boundaries.

	Exp 1 Ref.	Exp 4 ref.	
<i>achievement</i>	s376	s455	the act of accomplishing something
<i>arts</i>	s375	s267	studies intended to provide general knowledge and intellectual skills (rather than occupational or professional skills); "the college of arts and sciences"
<i>civilization</i>	s30	s693	a society in an advanced state of development
<i>culture</i>	s25	s292	a particular civilization at a particular stage
<i>effort</i>	s428	s318	a notable achievement: "the book was her finest effort"
<i>history</i>	s396	s287	the discipline that records and interprets past events involving human beings: "he teaches Medieval history"; "history takes the long view"
<i>society</i>	s55	s300	an extended social group having a distinctive cultural and economic organization

Figure 56 Interpretations of the words in paragraph 5.

Furthermore all the interpretations in this graph seem to me to be right on the spot, in order to represent the focus of the paragraph.

The graphs of this paragraph can as usual be found in the corresponding subsection of Appendix A-1.6.

7.4.7 Paragraph 6

The words of paragraph 6 are a bit complicated to relate to each other, because the concepts involved are somewhat remote. The system also fails to find them. The main reason is that the relationships involved are to some extent different from the ones recognised by the system. I will return to this in chapter 8.

7.4.8 Paragraph 7

The paragraph represents a stand-alone introduction to how the game can be completed and a very summarised mentioning of new content relative to older versions of the game. The paragraph is skidding across concepts and contexts at a hazardous speed. In fact there is almost a new context for each sentence in this paragraph. The skimming algorithm does however succeed to pinpoint two important concepts in this paragraph. Because of the interrelated polysemies of *course*, *score* and *way* the algorithm offers three slightly different interpretations of the paragraphs, all of which score the same - namely 4.

Haphazardly the *course* in the paragraph is part of the semantically almost empty expression *of course*, it just so happens that *course* relates meaningfully to K in several different ways in this paragraph. In fact none of the interpreted concepts for the word *course* are realized explicitly in the paragraph, whereas at least the s464 sense of the word is indeed present implicitly.

The polysemy of *score* relates to the distinction between the ACT_OF_SCORING and the resulting score itself. Both of these concepts are present in the paragraph, the ACT_OF_SCORING is however only present implicitly and the concept referred to in the text must be that of s468.

The proper meaning of *way* in this paragraph is clearly the COURSE OF CONDUCT-sense of the word, s195. The two different senses of *way* mentioned here represent the exact same distinction as the two different meaning of the word *course*. It is very interesting that related lexemes may be inheriting such distinctions from each other, but it will lead too far to elaborate here.

The *success-score* and *way-course* components do represent the HOW-TO-WIN context of the chapter pretty well I think. I have marked by an asterisk those senses of *course*, *score* and *success* that to me seems to express the paragraph the best. These senses are however not all present in the same interpretation, it will be interesting to see if this clears up when comparing paragraphs in the later experiments. With reference to the interpreting graphs present in table A-1.8 of appendix A-1, I would say that G2 offers the best interpretation of the paragraph but that the race is indeed very close between the three alternatives. See appendix A-4 for the glossary.

	Exp 1 Ref.	Exp 4 ref.	
course	s591	s458	general line of orientation: "the river takes a southern course"; "the northeastern trend of the coast"
	s597	s464*	a mode of action; "if you persist in that course you will surely fail"
game	s31	s684	the score needed to win a game; "he is serving for the game"
score	s609	s475	the act of scoring in a game or sport; "the winning score came with less than a minute left to play"
	s602	s468*	a number that expresses the accomplishment of a team or an individual in a game or contest; "the score was 7 to 0"
success	s569	s444	an attainment that is successful; "his success in the marathon was unexpected"; "his new play was a great success"
way	s339	s386	a line leading to a place or point: "he looked the other direction"; "didn't know the way home"
	s262	s195*	a course of conduct; "the path of virtue"; "we went our separate ways"; "our paths in life led us apart"; "genius usually follows a revolutionary path"

Figure 57 Interpretations of the words in the confusing paragraph 7.

7.4.9 Paragraph 8

This is the last paragraph that I included in the CIVIII experimental corpus so far. It represents quite a drastic shift in contextual focus. The main purpose of the paragraph is to help the reader benefit the most from the documentation of the game, rather than to appreciate the game experience itself. There are a lot of different concepts in this paragraph, mostly relating to different sections of the manual book and their respective contents.

Also present here is a blunder on my part, since the word *rear* is a verb in the paragraph, not a noun. Please recall that I made the noun lists by hand, after I demonstrated that a small computer program can be made to construct it. Obviously I chose the meaning of *rear* that first came to my mind, and that was the noun rather than the verb. I decided not to correct this mistake, however, since indeed the Tagger might as well have made the same mistake.

Not surprisingly the skimming algorithm identifies a relatively high number of concepts in this comparably long paragraph. There is a little confusion as to meaning of some of the words, namely *book*, *chapter* and *text*.

The polysemy of *book* relates to the distinction between, the CONTENT OF THE BOOK and the BOOK AS A PHYSICAL OBJECT on one side, and the STANDARD BOOK versus the part meaning of *book*, present in very large textual works like the Holy Bible and Lord of the Rings, that comprises several volumes/distinct bodies of text. The reading that I prefer for the paragraph in question is the s650 sense of the word.

The polysemy of *chapter* relates to the PART OF A BOOK meaning and the PART OF ORGANISATION reading already touched upon earlier in this chapter... Obviously the concept of *chapter* as a PART OF A BOOK is the one that fits the best with context in paragraph 8.

Finally, the word *text* is polysemous as it can both refer to text almost as a matter, like water or flour, and as a coherent clearly delimited portion of text. Clearly this is a very narrow distinction but I think the former of these alternatives (i.e.: s659) must be the one at play in the present context.

I have marked my preferred senses for the polysemous words of paragraph 8 by an asterisk in the table below.

The prototype suggests three slightly different interpretations for the paragraph the "best" of which scores 16 whereas the other two each score

15. The senses included in the highest scoring graph have been underlined in the table below. It can be seen that the algorithm actually picked out the same ones that I preferred - I think that other people would as well - for the three polysemous words.

There is a certain amount of noise present here as well. First there is the component that relates the verb/noun *rear* that obviously does not fit very well with the rest. Secondly it can be argued that the semantically almost empty component of *thing* and *change* would fit in every conceivable context and as such doesn't really help to identify any of them. On the whole I think the skimmer manages to get hold of most of the core concepts of the paragraph, and succeeds to present a context dominated by the different parts of some book of reference.

	Exp 1 Ref.	Exp 4 ref.	
<i>Binding</i>	<u>s777</u>	s637	the front and back covering of a book; "the book had a leather binding"
<i>Book</i>	<u>s790</u>	<u>s650*</u>	a copy of a written work or composition that has been published (printed on pages bound together); "I am reading a good book on economics"
	s791	s651	A major division of a long written composition; "the book of Isaiah")
	s794	s654	a book as a physical object: a number of pages bound together; "he used a large book as a doorstep"
<i>Change</i>	<u>s757</u>	<u>s617</u>	A thing that is different; "he inspected several changes before selecting one")
<i>Chapter</i>	<u>s615</u>	<u>s596*</u>	a subdivision of a written work; usually numbered and titled; "he read a chapter every night before falling asleep"
	s613	s594	A local branch of some fraternity or association; "he joined the Atlanta chapter"
<i>Head</i>	<u>s653</u>	<u>s517</u>	the subject matter at issue; "the question of disease merits serious discussion"; "under the head of minor Roman poets"
<i>Order</i>	s54	s299	a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today"
<i>Place</i>	<u>s712</u>	<u>s575</u>	the particular portion of space occupied by a physical object: "he put the lamp back in its place"
<i>Problem</i>	<u>s695</u>	<u>s559</u>	a question raised for consideration or solution; "our homework consisted of ten problems to solve"
<i>Rear</i>	<u>s632</u>	<u>s496</u>	the part of something that is furthest from the normal viewer: "he stood at the back of the stage"; "it was hidden in the rear of the store"
<i>reference</i>	<u>s805</u>	<u>s665</u>	a book to which you can refer for authoritative facts; "he contributed articles to the basic reference work on that topic"
<i>Text</i>	<u>s799</u>	<u>s659*</u>	the words of something written; "there were more than a thousand words of text"; "they handed out the printed text of the mayor's speech"; "he wants to reconstruct the original text"
	s797	s657	a book prepared for use in schools or colleges; "his economics textbook is in its tenth edition"
<i>Thing</i>	<u>s768</u>	<u>s628</u>	an entity that is not named specifically; "I couldn't tell what the thing was"

Figure 58 Paragraph 8 makes reference to a lot of different concepts.

The complete output of this paragraph can be found in appendix A-1 along with the glossary file of experiment 1. In appendix A-5 I included several tables. In particular the table "Interpretations relative to isolated paragraphs" illustrates what words get interpreted in what paragraph and the "Master scoring table for all experiments" compares the relative score of each experimental paragraph and some statistics as well. See appendix A-4 for the glossary.

7.4.10 Conclusion on Experiment 1

I think it is fair to conclude that the skimming algorithm does catch most of the core concepts of reasonably well-presented paragraphs of the corpus. It fails, to some extent when handling short summary-like paragraphs involving several distinct contexts. If the algorithm was to be modified to handle implicit concepts as well.

By “implicit concept” here, I refer to a concept that to a human is clearly present *between the lines* of the textual data but is *not realized* by any actual word, i.e.: *no word in the text has that concept among its possible meanings*. Referring to section 7.4.2 no mentioning of a concept like GEOGRAPHICAL_TERRAIN distinct from MENTAL_TERRAIN occurs in paragraph 1 while it clearly present to a human reader via the concepts explicitly realized by *terrain, mountain, river* etc.

Allowing edges between concepts that are explicitly realized in the text, via ones that are not, would probably improve a lot on the interpretational accuracy. I have already mentioned the possible consequences for degree of noise and computational complexity from such a modification. (see also the conclusion in section 7.8).

7.5 Experiment 2 - first and subsequent paragraphs in pairs.

How do changes in context manifest themselves ? This is the core question intertwining all the issues of part II of the experimental interests mentioned in section 7.3 of this chapter. To continue the reasoning of the rest of this work, changes in context could be marked by several incidents. A word that has one meaning up to a certain point, and then changes meaning for a while, could for instance be an indication that the context has shifted. On the other hand it could just mean that system finally got the intended meaning right and the context is still the same as before. Similarly lexemes that all of the sudden become contradictory to the interpretation of other words in a particular sequence, could be a sign of contextual shift.

The prototype can at present not decide on such questions but uniformly treats each paragraph, i.e.: a sequence of words preceded by a “p.”, as its own context. It is however very simple to experiment with expanded paragraphs on a small scale. All that is required is to remove selected paragraph markers between paragraphs. The prototype will then attempt to interpret the combined sequence as one paragraph. Comparing to the interpretations of the isolated paragraph will then perhaps help pinpointing the conditions for deciding on the question of continuation of previous context or distinction from it.

Experiment 2 is the first in a row of gradually larger such combined sequences. Referring to the source text of Figure 52, it looks like paragraph 0 is a kind of super paragraph to paragraphs 1 through 5, while paragraphs 6, 7 and 8 are intended as stand-alone paragraphs – in particular distinct from the first six paragraphs. Experiment 2 combines paragraph 0 with each of the other paragraphs in turn, to see if paragraph 0 is closer related to paragraphs 1 through 5 than to paragraphs 6, 7 and 8 respectively. The resulting interpretations are then compared to the interpretations of the respective paragraphs in isolation as established in the previous section.

7.5.1 Paragraphs 0 and 1 as one.

When joining together two paragraphs like this, the “neutral” result could be expected to be the graphs of the respective sequences in one so to speak. Referring to sections 7.4.1 and 7.4.2 the system offered a single interpreting graph for each of the paragraphs. The combined sequence result in two alternative interpreting graphs, that both score the same namely 18, slightly better than the sum of the isolated scores. One of these is just the “combination” of graphs with no surprises. The other one offers a new interpretation of *force*, namely s310 that related to the ORGANISATIONAL_UNIT of s672. While the military *force* is certainly not the one referred to in paragraph 0, the resulting component does subjectively represent the context of the combined sequence very well. The ORGANISATIONAL_UNIT proposed in G2, seems spot on to me.

All in all it is indeed hard to choose between the two interpretations offered, and I think that scoring system performs well in expressing this

doubt by holding on to both for now. However the interpretations of *area* /*territory* and *force* /*impulse* remain misleading and dubious, respectively. Only the issues relating to *force* already mentioned, indicates a possible conflict of interpretations between the paragraphs. But at least they can be seen to have overlapping contexts.

	Exp 2 Ref.	Exp 4 ref.	
Area	s80	s57	a subject of study; "it was his area of specialization"; "areas of interest include..."
<i>civilization</i>	s25	s693	a society in an advanced state of development
<i>Culture</i>	S88	s292	a particular civilization at a particular stage
Force	S20	S312	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
	S18	s310	a unit that is part of some military service; "he sent Caesar a force of six thousand men"
<i>Goal</i>	s4	s4	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
Impulse	S29	s19	an impelling force or strength; "the car's momentum carried it off the road"
<i>Objective</i>	S23	s14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>Society</i>	S42	s300	an extended social group having a distinctive cultural and economic organization
Territory	S66	s44	an area of knowledge or interest; "his questions covered a lot of territory"
<i>Unit</i>	S96	s672	an organization regarded as part of a larger social group; "the coach said the offensive unit did a good job"; "after the battle the soldier had trouble rejoining his unit"

Figure 59 The joined sequence of paragraph 0 and 1 caused the skimmer to consider a new meaning for the word *force*. Also *Unit* is a newcomer among the interpreted words stemming from the combination of the two paragraphs.

Figure 60 summarises the results of the isolated sequences and compares to the result of the combined sequence. For the upcoming sequences I will not show this table in the main text but refer to the respective appendix, Appendix A-5. For more details on the results of this sequence see Appendix A-2.1. See appendix A-4 for the glossary.

	dall-p0	dall-p1	d0X-0	d0X-0
Interpretation	#1	#1	#1	#2
Score	15	2	18	18
Token				
<i>Area</i>	-----	s57	s57	s57
<i>Civilization</i>	s693	-----	s693	s693
<i>Culture</i>	s292	-----	s292	s292
<i>Force</i>	s312	-----	s312	s310
<i>Goal</i>	s4	-----	s4	s4
<i>Impulse</i>	S19	-----	s19	-----
<i>Objective</i>	S14	-----	s14	s14
<i>Society</i>	s300	-----	s300	s300
<i>Territory</i>	-----	s44	s44	s44
<i>Unit</i>	-----	-----	-----	s672

Figure 60 Interpretations of the two paragraphs in isolation compared to the interpretations of the combined sequence. Changes relative to the earlier interpretations have been boldfaced. Normal types indicates a continuation with respect to earlier results and "-----" in an interpretational column for a word, indicates that the word did not have related meanings in that interpretation. Sense numbers refers to the legend of Experiment 4 (see appendix A.1-4) See also Appendix 5.

7.5.2 Paragraphs 0 and 2 as one.

	Exp 2 Ref.	Exp 4 ref.	
<i>civilization</i>	s25	s693	a society in an advanced state of development
<i>culture</i>	s88	s292	a particular civilization at a particular stage
<i>force</i>	s20	s312	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
<i>goal</i>	s4	s4	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
<i>impulse</i>	s29	s19	an impelling force or strength; "the car's momentum carried it off the road"
<i>objective</i>	s23	s14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>power</i>	s251	s145	possession of the qualities (especially mental qualities) required to do something or get something done; "danger heightened his powers of discrimination"
<i>purpose</i>	s200	s109	an anticipated outcome that is intended or guides your planned actions; "his intent was to provide a new translation"; "it was created with the conscious aim of answering immediate needs"; "he made no secret of his designs"
<i>rate</i>	s197	s107	(British) a local tax on property (usually used in the plural)
	s198	s108	the relative speed of progress or change; "he lived at a fast pace"; "he works at a great rate"; "the pace of events accelerated"
	s195	s105	a magnitude or frequency relative to a time unit; "they traveled at a rate of 55 miles per hour"; "the rate of change was faster than expected"
<i>science</i>	s174	s181	ability to produce solutions in some problem domain; "the skill of a well-trained boxer"; "the science of pugilism"
<i>society</i>	s42	s300	an extended social group having a distinctive cultural and economic organization
<i>speed</i>	s205	s114	a rate (usually rapid) at which something happens; "the project advanced with gratifying speed"
	s203	s112	distance travelled per unit time
<i>tax</i>	s212	S178	charge against a citizen's person or property or activity for the support of government

Figure 61 The context of the combined sequences of paragraphs 0 and 2 perfectly continues the contexts of both and even elaborates on it by adding a vertex to an already established component.

This sequence combines the sequences originating from the paragraphs named “Five Impulses ...” and “Economics” in Figure 52. The skimming algorithm offers three slightly different interpretations of the joined sequence one of which scores 24 and the remaining two scores 23. This represents quite an increase in score when comparing to the sum of the isolated scores, namely at most 20. All the components of the respective best scoring isolated interpretations persist in the best scoring interpretation of the combined sequence. This indicates that there are no conflicts of interpretations at all. That the two paragraphs have almost identical contexts is emphasised by the addition of the *purpose* vertex to the goal-objective component of paragraph 0. Also the word *civilization* is mentioned in both paragraphs with no contradictions in interpretation¹⁰. The combination of these two sequences serves as an excellent example, I think, of a continuation of contexts.

Still, however, the interpretations offered for the words *force*, *impulse*, *power* and *rate* remain problematic. In particular, the scoring scheme persists in choosing the specialised but wrong interpretation for *rate*, because of the close relationship to *tax*. In spite of this, the best scoring

¹⁰ (- Apart from the curiosity that *civilization* refers to both the name of the game and to the cultural organisation of people in paragraph 0 and other places as well. The prototype is not equipped with tools to accommodate for inventive word use, where a known expression is assigned new meaning, specific only to the one context at hand, as it is.)

interpretation (the underlined senses in the table below) does quite adequately represent the common subjective context of the sequences.

Find the complete graphs and statistics for this experiment in appendices A-2.2 and A-5. See appendix A-4 for the glossary.

7.5.3 Paragraphs 0 and 3 as one.

	Exp 2 Ref.	Exp 4 ref.	
<i>civilization</i>	<u>s25</u>	<u>s693</u>	a society in an advanced state of development
<i>culture</i>	<u>s88</u>	<u>s292</u>	a particular civilization at a particular stage
<i>force</i>	<u>s20</u>	<u>s312</u>	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
	s18	s310	a unit that is part of some military service; "he sent Caesar a force of six thousand men"
<i>goal</i>	<u>s4</u>	<u>s4</u>	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
<i>impulse</i>	<u>s42</u>	<u>s19</u>	an impelling force or strength; "the car's momentum carried it off the road"
<i>objective</i>	<u>s23</u>	<u>s14</u>	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>society</i>	<u>s55</u>	<u>s300</u>	an extended social group having a distinctive cultural and economic organization
<i>unit</i>	s96	s672	an organization regarded as part of a larger social group; "the coach said the offensive unit did a good job"; "after the battle the soldier had trouble rejoining his unit"

Figure 62 The interpretations of the combined sequences of paragraphs 0 and 3. Note that this sequence raises doubt with respect to the meaning of “force”. Indeed the notion of “force” differs in the two paragraphs even if the notion of “military force” is only implicitly present via “unit” in the context of paragraph 3.

The combination of paragraphs 0 and 3 is almost trivial. In spite of the fact, however, that the skimming algorithm failed to find any contextual components in paragraph 3, when it was treated in isolation it now finds two alternative interpretations of the combined sequence. One scores 15 equal to the interpretation of paragraph 0. The other scores only 14, which constitutes some contrastive contexts.

The contrast is that the concept of *force* in paragraph 0 is different from the one that would make sense in paragraph 3. As already pointed out the *force* of paragraph 0 is a PSYCHOLOGICAL FORCE OR URGE and even though it gets interpreted in that paragraph as a PHYSICAL FORCE it certainly is not the MILITARY FORCE that relates to concepts realized by *unit* in paragraph 3.

It is a valid argument that the word *force* is not even mentioned in paragraph 3 and so it should not prevent the proper interpretation of *unit* even if it means something else in paragraph 0. And rightly so. The combination of sequences that I conduct in these experiments are just that, purely experimental. They are done to see if contextual overlap and contrasts can be identified through the mechanisms of the skimming prototype.

In this respect, when one combines two paragraphs - one of which has one interpretation and one which has none (i.e.: having an interpreting graph that scores 0, having no components) - and have as result two contradicting interpretations, it is perhaps cause to investigate further.

In this case the alternative interpretation is not due to failure of the skimming algorithm but the consequence of actual and very subtle differences in the contexts of the two paragraphs. It is worth to note when such subtle differences can be identified by the computer program.

See the Appendices A-2.3 and A-5 for the graphs and statistics of the sequence. See appendix A-4 for the glossary.

7.5.4 Paragraphs 0 and 4 as one.

	Exp 2 Ref.	Exp 4 ref.	
<i>aggression</i>	<u>s471</u>	<u>s245</u>	violent action that is hostile and usually unprovoked
<i>civilization</i>	<u>s25</u>	<u>s693</u>	a society in an advanced state of development
<i>culture</i>	<u>s88</u>	<u>s292</u>	a particular civilization at a particular stage
	s90	s294	the tastes in art and manners that are favored by a social group
<i>force</i>	s20	s312	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
	<u>s22</u>	<u>s314</u>	an act of aggression (as one against a person who resists); "he may accomplish by craft in the long run what he cannot do by force and violence in the short one"
<i>goal</i>	<u>s4</u>	<u>s4</u>	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
<i>impulse</i>	s42	s19	an impelling force or strength; "the car's momentum carried it off the road"
<i>objective</i>	<u>s23</u>	<u>s14</u>	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>society</i>	<u>s55</u>	<u>s300</u>	an extended social group having a distinctive cultural and economic organization
<i>taste</i>	s437	s211	delicate discrimination (especially of aesthetic values); "arrogance and lack of taste contributed to his rapid success"; "to ask at that particular time was the ultimate in bad taste"

Figure 63 Note the interesting reinterpretation of force and also the alternate component of “culture-taste” that was considered but didn’t make it.

The sequence resulting from combining the sequences of paragraphs 0 and 4 also holds a few surprises. The skimming algorithm finds only one interpretation that qualifies through its score for output but examining the possibly connected words of the sequence it is apparent that it has been considering other interpretations as well, others that were found too light. The interpretation that was output scores 18, which represents and substantial increase with respect to the combined score of the isolated interpretations, i.e.: $15 + 0 = 15$.

We see that *aggression* got interpreted as related to *force* and indeed FORCEFUL AGGRESSION is an important concept in paragraph 4 named “Conquest” in Figure 52. In the interpretation of paragraph 0 *force* was assigned s312 and that apparently was also considered here but an interpretation involving that sense of *force* would not interpret *aggression* and thus would interpret fewer lexemes than the chosen one.

Also *culture* has been considered for a new meaning namely s294 relating to a meaning of *taste*. If *culture* was to be assigned this meaning it would have to be removed from the strong *society*, *civilization* component and the loss in score was obviously too great.

Subjectively; I appreciate very much the choices made by the algorithm in this sequence. I think the inclusion of *aggression* adds significantly and importantly to the context even if the urge-like *force* and *impulse* of paragraph 0 had to go. While the interpretation of *taste* must be what the author had in mind it is not nearly as important to context as the strong constellation of *society-civilization-culture*. So the skimming

algorithm all in all performs very well in this sequence, though it may of course be a “lucky punch”, so to speak.

See the actual graph and statistics in appendices A-2.4 and A-5. See appendix A-4 for the glossary.

7.5.5 Paragraphs 0 and 5 as one.

	Exp 2 Ref.	Exp 4 ref.	
<i>achievement</i>	§553	§455	the act of accomplishing something
<i>arts</i>	§552	§267	studies intended to provide general knowledge and intellectual skills (rather than occupational or professional skills); "the college of arts and sciences"
<i>civilization</i>	§25	§693	a society in an advanced state of development
<i>culture</i>	§88	§292	a particular civilization at a particular stage
<i>effort</i>	§605	§318	a notable achievement: "the book was her finest effort"
<i>force</i>	§20	§312	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
<i>goal</i>	§4	§4	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
<i>greatness</i>	§613	§326	the property possessed by something or someone of outstanding importance
<i>history</i>	§573	§287	the discipline that records and interprets past events involving human beings: "he teaches Medieval history"; "history takes the long view"
<i>importance</i>	§32	§22	the quality of being important and worthy of note; "the importance of a well-balanced diet"
<i>impulse</i>	§29	§19	an impelling force or strength; "the car's momentum carried it off the road"
<i>objective</i>	§23	§14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>Society</i>	§42	§300	an extended social group having a distinctive cultural and economic organization
<i>Value</i>	§558	§272	the quality (positive or negative) that renders something desirable or valuable; "the Shakespearean Shylock is of dubious value in the modern world"

Figure 64 The interpretation of the words in paragraph 0 and 5 in combination provoke a new important component in their common context. Sadly, the mildly misleading interpretations of force and impulse persist also in this sequence.

The combination of paragraphs 0 and 5 results in an interpretation that is a perfect continuation of the contexts of the isolated paragraphs. There are no contradicting interpretations at all. The single interpretation of this sequence scores a nice 31 compared to the 26 that is the sum of the scores of the paragraphs interpreted in isolation. The word *importance* of paragraph 0 causes the words *greatness* and *value* of paragraph 5 to be connected. This new component rightly assumes a prominent place in the interpretation and is itself responsible for 4 of the 5 points difference in score. The last point stems from the reappearance of *force* in paragraph 5. In paragraph 5 there is subjectively no doubt that *force* should have been interpreted as a MILITARY FORCE rather than the erroneous PHYSICAL FORCE inherited from paragraph 0 where it didn't really sit right either.

Apart from this unfortunate problem, the interpretations offered by the skimmer for this sequence very well illustrates the concepts clearly intended in the original paragraphs – concepts, that also represents an almost full cover of the core focal points of the involved contexts. It is true however, that the notion of IMPORTANCE implicitly present in paragraph 5 is different from the one intended in paragraph 0. The interpretations of *value* and in particular of *greatness* could not have been made without IMPORTANCE or a similar concept in the sequence since they are not directly related. In a way, this illustrates what might be gained from allowing the system to consider

intermediate/shadow concepts when checking for relations between concepts. If this had been allowed, the *value – greatness* component would have been established already when interpreting paragraph 5 in isolation. On the other hand, a lot of “noisy”/erroneous interpretations would perhaps also have emerged. It seems worth doing experiments in this direction but it will have to wait till another opportunity as it would expand the framework of this project into undesired proportions.

In all, I think it is fair to that the skimmer performs very acceptably in interpreting this sequence.

The graph and statistics of this sequence can be found in the usual appendices.(A-2.5, A-5). See appendix A-4 for the glossary.

7.5.6 Paragraphs 0 and 6 as one.

The context of the words of paragraphs 0 and 6 interpreted as one paragraph is an exact copy of the context presented for paragraph 0 in isolation. The fact, that there are no contradictions and indeed that the only change resulting from the combination of the sequences is a +2 increase in score, is perhaps the most clear example of a complete continuation of the context established in paragraph 0. The difference in score stems from the reappearance of *civilization* in paragraph 6. The relations to both *society* and *culture* cause the entire difference in score of the component and entire interpretation as well.

This is the first combination of sequences that does not add anything that substantiates a common context. Sure, *civilization* gets mentioned again but there are no new interpretations gained from the combination. On the other hand, there is nothing that prevents a common context either.

The graph and statistics of the sequence are included in the appendices (A-2.6, A-5). See appendix A-4 for the glossary.

	Exp 2 Ref.	Exp 4 ref.	
<i>civilization</i>	s25	s693	a society in an advanced state of development
<i>Culture</i>	s88	s292	a particular civilization at a particular stage
<i>Force</i>	s20	s312	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
<i>goal</i>	s4	s4	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
<i>impulse</i>	s29	s19	an impelling force or strength; "the car's momentum carried it off the road"
<i>objective</i>	s23	s14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>Society</i>	S42	s300	an extended social group having a distinctive cultural and economic organization

Figure 65 The interpretations offered for the combined sequence of words form paragraphs 0 and 6 are identical to those offered for the words of paragraph 0 in isolation. It scores a little higher because “civilization” appears in both paragraphs and thus continues to be an important concept in the context.

7.5.7 Paragraphs 0 and 7 as one.

The skimmer offers only two interpretations of the combined sequence of paragraphs 0 and 7 as opposed to three for the sequence of paragraph 7. The two interpretations both score 21 representing a slight increase in score when compared to the sum of scores of the individual paragraphs, 19.

	Exp 2 Ref.	Exp 4 ref.	
<i>chapter</i>	s894	s594	a local branch of some fraternity or association; "he joined the Atlanta chapter"
<i>civilization</i>	s25	s693	a society in an advanced state of development
<i>course</i>	s868	s458	general line of orientation: "the river takes a southern course"; "the northeastern trend of the coast"
	s874	s464	a mode of action; "if you persist in that course you will surely fail"
<i>culture</i>	s88	s292	a particular civilization at a particular stage
<i>force</i>	s20	s312	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
<i>game</i>	s46	s684	the score needed to win a game; "he is serving for the game"
<i>goal</i>	s4	s4	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
	s6	s6	a successful attempt at scoring; "the winning goal came with less than a minute left to play"
<i>impulse</i>	s29	s19	an impelling force or strength; "the car's momentum carried it off the road"
<i>objective</i>	s23	s14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>score</i>	s886	s475	the act of scoring in a game or sport; "the winning score came with less than a minute left to play"
	s879	s468	a number that expresses the accomplishment of a team or an individual in a game or contest; "the score was 7 to 0"
<i>society</i>	s42	s300	an extended social group having a distinctive cultural and economic organization
	s41	s299	a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today"
<i>success</i>	s846	s444	an attainment that is successful; "his success in the marathon was unexpected"; "his new play was a great success"
<i>way</i>	s480	s386	a line leading to a place or point: "he looked the other direction"; "didn't know the way home"
	s357	s195	a course of conduct; "the path of virtue"; "we went our separate ways"; "our paths in life led us apart"; "genius usually follows a revolutionary path"

Figure 66 Goal changes its meaning and also society were considered for a disturbing interpretation when interpreting paragraphs 0 and 7 as one.

The main differences from the interpretations of the paragraphs in isolation concern themselves with the interpretation of *goal* as related to the *score-success* component of paragraph 7, namely via s6, where it has up until now been interpreted as related to *objective* via s4. I find that to be the wrong interpretation of the word. This is because the words *goal* and *score* are so closely related in contexts pertaining to sports. In this sequence both *score* and *goal* get interpreted to misleading meanings but especially that *goal* changes meaning should be an important sign of something special happening when combining the sequences. Another such indication is that the strong concept realized by *society* clearly has been considered for replacement as well, namely *society* as related to a questionable meaning of *chapter* as an ORGANISATIONAL BRANCH OF A (ESOTERIC) SOCIETY.

Finally the fact that *objective* is excluded from both interpretations should also be considered carefully when deciding whether the sequences are better interpreted isolated or in combination.

There is a slight increase in score, but I still think there are several differences in behaviour when compared to the earlier combined sequences. Note that it is not trivially true that the score of the combined sequences of two paragraphs is simply the sum of their respective scores. New words - having new meanings - may facilitate relationships to be formed that could not be formed in the isolated paragraphs on their own.

See the appendices A-2.7 and A-5 for graphs and statistics of this sequence. See appendix A-4 for the glossary.

7.5.8 Paragraphs 0 and 8 as one.

The combined sequences of paragraphs 0 and 8 have only one interpretation coming out on top scoring 37 compared to 31 that results from adding the scores of best interpretations of the individual paragraphs.

The sequence introduces no new words to be interpreted. Words *text* and *thing* get reinterpreted and words *chapter* and *order* lose their contextual relations.

The word *text* is now interpreted as a kind of *book* just like *reference*. The interpretation of *chapter* is lost this way and that of *order* with it.

The word *thing* changes from an ENTITY to an OBJECTIVE and combines with *objective* and *goal*. While not important to the contextual focus of either paragraph, the meaning of *thing* intended in paragraph 7 is quite clearly the ENTITY reading.

	Exp 2 Ref.	Exp 4 Ref.	
<i>binding</i>	s1103	s637	the front and back covering of a book; "the book had a leather binding"
<i>book</i>	s1116	s650	a copy of a written work or composition that has been published (printed on pages bound together); "I am reading a good book on economics"
	s1117	s651	A major division of a long written composition; "the book of Isaiah"
	s1120	s654	a book as a physical object: a number of pages bound together; "he used a large book as a doorstep"
<i>change</i>	s1083	s617	A thing that is different; "he inspected several changes before selecting one"
<i>chapter</i>	s896	s596	a subdivision of a written work; usually numbered and titled; "he read a chapter every night before falling asleep"
	s894	s594	A local branch of some fraternity or association; "he joined the Atlanta chapter"
<i>civilization</i>	s25	s693	a society in an advanced state of development
<i>culture</i>	s88	s292	a particular civilization at a particular stage
<i>force</i>	s20	s312	physical energy or intensity; "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
	s18	S310	a unit that is part of some military service; "he sent Caesar a force of six thousand men"
<i>goal</i>	s4	s4	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
<i>head</i>	s980	s517	the subject matter at issue; "the question of disease merits serious discussion"; "under the head of minor Roman poets"
<i>impulse</i>	s29	s19	an impelling force or strength; "the car's momentum carried it off the road"
<i>objective</i>	s23	s14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>order</i>	s41	s299	a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today"
<i>place</i>	s1039	s575	the particular portion of space occupied by a physical object: "he put the lamp back in its place"
<i>problem</i>	s1022	s559	a question raised for consideration or solution; "our homework consisted of ten problems to solve"
<i>rear</i>	s959	s496	the part of something that is furthest from the normal viewer: "he stood at the back of the stage"; "it was hidden in the rear of the store"
<i>reference</i>	s1131	s665	a book to which you can refer for authoritative facts; "he contributed articles to the basic reference work on that topic"
<i>society</i>	s42	s300	an extended social group having a distinctive cultural and economic organization
	s41	s299	a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today")
<i>text</i>	s1125	s659	the words of something written; "there were more than a thousand words of text"; "they handed out the printed text of the mayor's speech"; "he wants to reconstruct the original text"
	s1123	s657	a book prepared for use in schools or colleges; "his economics textbook is in its tenth edition"
<i>thing</i>	s1090	s624	a special objective: "the thing is to stay in bounds"
	s1094	s628	an entity that is not named specifically; "I couldn't tell what the thing was"

<i>unit</i>	<u>s96</u>	<u>s672</u>	an organization regarded as part of a larger social group; "the coach said the offensive unit did a good job"; "after the battle the soldier had trouble rejoining his unit"
-------------	------------	-------------	--

Figure 67 No new words introduced, two words reinterpreted and another two losing contextual relations. Still, however, a +6 increase in score.

Subjectively the two paragraphs perhaps represent the two most different contexts in the small corpus. Paragraph 0 makes introductions to some key concepts of the game while paragraph 8 is kind of a meta text about how the manual is organised, how to use it and also what informative help is available within the computer game itself. It is still reasonably clear that the two paragraphs do indeed share contexts to some extent.

Even though the objective signs are not as convincing as was the case in section 7.5.7, there are indications that the paragraphs might be better interpreted in isolation - as there clearly are lexemes in both, that won't go together. Again, the indications are very subtle and regrettably not immediately visible in the scoring of the sequences. It looks like the signals for contextual shift have to be found by keeping track of how the words get interpreted, reinterpreted and demoted from relation to the respective contexts.

The graph and statistics for this last sequence of Experiment 2 is found in the appendices A-2.8 and A-5. See appendix A-4 for the glossary.

7.5.9 Conclusions of Experiment 2

The purpose of Experiment was twofold. First I wanted to examine how contextual shifts might manifest themselves and secondly I wanted to how the sketchy contexts of experiment 1 might be refined with slightly larger sequences to draw information from.

Combining the sequence of paragraph 0 with any of the paragraphs 1 through 5 results in a relative increase in the scores of the interpretations while no lexemes are demoted with respect to their isolated interpretations. Furthermore almost all of the combined sequences actually add important concepts to the perceived contexts.

Combining paragraph 0 with paragraphs 6, 7 or 8 also increases the relative score, but the number of added concepts to the context is relatively lower and the additions seem more accidental than anything else. These combinations do result in the demotion of certain lexemes from the respective isolated interpretations as well.

Consequently I think it is fair to say that there are signs to look for when trying to establish differences in contextual focus but the scoring, as it is, is not such a sign.

Appendix A-5 holds a master scoring table for all experiments where the scores of the individual paragraphs and also the combined sequences can be compared. Each of the combined sequences also has a table in A-5 comparing their interpretations and scores relative to the respective individual paragraphs.

7.6 Experiment 3 - paragraphs 0-5 as one - 6, 7 and 8 isolated.

Combining the sequences of the first 6 paragraphs in order to interpret them as one should make it clear whether the skimmer has any ability to decide if they have overlapping contexts. In Figure 52 it is subjectively clear that indeed paragraph 0 serves as an introduction to the five paragraphs to follow – one for each “impulse of civilization” as the chapter headline suggests. The purpose of this experiment is to see if this structure is at all visible in the results of the skimmer.

To this end, the interpretations and contexts of the combined paragraphs 0 through 5 should be compared to those of earlier experiments.

It should be noted that only one interpreting graph is deemed “good” enough for being the “best”, so to speak. It recognises 31 lexemes out of 93 possible corresponding to 49 instances out of 125. It scores 61 compared to only 33 (- the sum of scores corresponding to the individual paragraphs).

7.6.1 New lexemes resulting from experiment 3

The combined sequences recognises a large number of lexemes that were not acknowledged before.

Aggression is interpreted as relating to *force* (- *force* itself gets reinterpreted and changes meaning relative to its interpretation in par.0), and while the corresponding meaning of *force* is not the kind intended in paragraph 0, it is certainly the kind of *aggression* intended in paragraph 4. This way it seems fair to say that the recognised kind of *force* is certainly present in the corpus even if only implicitly and its inclusion is not in any way misleading as to the context of the corpus text.

The same could be said about the word *darkness* (paragraph 1) is interpreted as relating to the a meaning of *value* (paragraph 5). The subjectively preferred interpretation of *value* is quite clearly s272 instead of the somewhat troublesome s273 that relates to the meaning of *darkness*.

The word *discovery* of paragraph 4 is finally interpreted as an ACHIEVEMENT via EFFORT – both in paragraph 5. This component now strongly represents a core element of the CivilizationIII context.

Another new concept is that of *greatness* (par. 5) interpreted in relation to the meaning of *importance* (par. 0). While this component was already recognised in the combination of pars. 0 and 5, it remains new when comparing to the isolated paragraphs. Achieving the kind of GREATNESS as intended in par. 5 is certainly on of the IMPORTANT THINGS referred to in par. 0.

Then we get an interpretation of *knowledge* (par. 3) neatly relating to POWER and SCIENCE (pars. 2 and 3 respectively).

The words *path* and *way* are interpreted in relation to each others meanings. Surely these words are in relation, but in another aspect than the one indicated by their meanings. They are interpreted as ordinary TRAVEL ROUTES where they clearly refer to PATHS TO SUCCESS in the corpus text.

The words *people*, *population* and *world* are assigned interpretations that are related to each other and together they form a strong and important component that is central to the content of the game manual.

The same is true for the words *purpose*, *goal* and *objective* that get very accurate interpretations with regard to the subjective corpus context.

While not nearly as important the words *range* and *variety* get pretty accurately interpreted as well.

7.6.2 Lexemes that were removed from contextual relation because of the concatenation of paragraphs.

The words *impulse*, *taste*, *tax* and *unit* were all interpreted in one or more of the previous experiments, but can not be made to relate to the context of the interpretation of this experiment. Two of these words were interpreted dubiously in earlier experiments (*impulse* and *tax* - both in paragraph 0). The word *taste* got interpreted well enough but the corresponding concept was - and is - not very important to the context, so the loss is not a big one. The word *unit* does however represent a very important concept to the context of the game. When playing the game, the player spends a lot of time, researching, building and manipulating his units. So why did *unit* have to go uninterpreted then? Quite simply because the respective meaning of the word is related to the MILITARY ORGANISATIONAL FORCE and not to the VIOLENCE-LIKE FORCE that relates to AGGRESSION in this interpretation. So we have three different concepts in play that can all be referred to by the word *force*. The PSYCHOLOGICAL URGE-LIKE FORCE of paragraph 0, the EXPEDITIONARY FORCE that relates to unit and finally the kind of force that can also be referred to using the word *violence* and relates to AGGRESSION. Of these three only the first is explicitly present in the context, expressively referred by use of the word *force*, while the last two are only indirectly present via their respective relations to words like *unit* and *aggression*.

I have restricted the skimmer to disallow more than one meaning for each word in a given paragraph and thus only one of the three meanings can figure in one particular interpretation, there are no words in the text that can refer to these concepts expressively. Perhaps this pertains to the suggested inclusion of shadow concepts, that does not need expressive realization in the text but bridges between concepts that are expressively represented.

7.6.3 Lexemes that persist in the interpretation of the combined sequences.

Finally all but four lexemes that figured in one or more interpretations of the individual sequences, are also related when the sequences are interpreted in combination.

While this seems perhaps not so interesting, it does suggest that these lexemes are not in semantic conflict with each other. In this way the context of the combined sequences serves a very nice continuation of all the contexts developed in the respective isolated lexemes. Only four lexemes would benefit from interpreting the paragraphs individually. This has to be compared to 27 lexemes that either “couldn’t care less” or actually benefits from the combined interpretation, even though some of these should be

counted as noise (- at least two of the complaining lexemes are kind of noisy themselves.)

	Exp 3 Ref.	Exp 4 Ref.	
<i>achievement</i>	s313	s455	the act of accomplishing something
<i>aggression</i>	s269	s245	violent action that is hostile and usually unprovoked
<i>area</i>	s57	s57	a subject of study; "it was his area of specialization"; "areas of interest include..."
<i>arts</i>	s312	s267	studies intended to provide general knowledge and intellectual skills (rather than occupational or professional skills); "the college of arts and sciences"
<i>civilization</i>	s374	s693	a society in an advanced state of development
<i>culture</i>	s340	s294	the tastes in art and manners that are favored by a social group
	s338	s292	a particular civilization at a particular stage
<i>darkness</i>	s40	s40	having a dark or somber color
<i>discovery</i>	s214	s198	the act of discovering something
<i>effort</i>	s364	s318	a notable achievement: "the book was her finest effort"
<i>force</i>	s360	s314	an act of aggression (as one against a person who resists); "he may accomplish by craft in the long run what he cannot do by force and violence in the short one"
	s358	s312	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
	s356	s310	a unit that is part of some military service; "he sent Caesar a force of six thousand men"
<i>frequency</i>	s188	s182	the number of occurrences within a given time period (usually 1 second); "the frequency of modulation was 40 cycles per second"
<i>goal</i>	s4	s4	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
<i>greatness</i>	s372	s326	the property possessed by something or someone of outstanding importance
<i>history</i>	s333	s287	the discipline that records and interprets past events involving human beings: "he teaches Medieval history"; "history takes the long view"
	s332	s286	all that is remembered of the past as preserved in writing; a body of knowledge: "the dawn of recorded history"; "from the beginning of history"
<i>importance</i>	s22	s22	the quality of being important and worthy of note; "the importance of a well-balanced diet"
<i>impulse</i>	s19	s19	an impelling force or strength; "the car's momentum carried it off the road"
<i>knowledge</i>	s168	s362	the psychological result of perception and learning and reasoning
<i>objective</i>	s14	s14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
<i>path</i>	s200	s194	a way especially designed for a particular use
	s199	s193	an established line of travel or access
<i>people</i>	s337	s291	(plural) any group of human beings (men or women or children) collectively; "old people"; "there were at least 200 people in the audience"
<i>population</i>	s192	s186	the people who inhabit a territory or state; "the population seemed to be well fed and clothed"
<i>power</i>	s150	s145	possession of the qualities (especially mental qualities) required to do something or get something done; "danger heightened his powers of discrimination"
<i>purpose</i>	s110	s109	an anticipated outcome that is intended or guides your planned actions; "his intent was to provide a new translation"; "it was created with the conscious aim of answering immediate needs"; "he made no secret of his designs"
<i>range</i>	s247	s223	a variety of different things or activities; "he answered a range of questions"; "he was impressed by the range and diversity of the collection"
<i>rate</i>	s108	s108	the relative speed of progress or change; "he lived at a fast pace"; "he works at a great rate"; "the pace of events accelerated"
	s107	s107	(British) a local tax on property (usually used in the plural)
	s105	s105	a magnitude or frequency relative to a time unit; "they traveled at a rate of 55 miles per hour"; "the rate of change was faster than expected"
<i>science</i>	s187	s181	ability to produce solutions in some problem domain; "the skill of a well-trained boxer"; "the science of pugilism"
<i>society</i>	s346	s300	an extended social group having a distinctive cultural and economic organization
<i>speed</i>	s115	s114	a rate (usually rapid) at which something happens; "the project advanced with gratifying speed"
	s113	s112	distance travelled per unit time
<i>taste</i>	s235	s211	delicate discrimination (especially of aesthetic values); "arrogance and lack of taste contributed to his rapid success"; "to ask at that particular time was the ultimate in bad taste"
<i>tax</i>	s184	s178	charge against a citizen's person or property or activity for the support of government
<i>territory</i>	s44	s44	an area of knowledge or interest; "his questions covered a lot of territory"
<i>unit</i>	s205	s672	an organization regarded as part of a larger social group; "the coach said the offensive unit did a good job"; "after the battle the soldier had trouble rejoining his unit"
<i>value</i>	s319	s273	relative darkness or lightness of a color: "I establish the colors and principal values by organizing the painting into three values--dark, medium...and light"-Joe Hing Lowe
	s318	s272	the quality (positive or negative) that renders something desirable or valuable; "the Shakespearean Shylock is of dubious value in the modern world"

<i>variety</i>	s89	s89	a collection containing a variety of sorts of things; "a great assortment of cars was on display"; "he had a variety of disorders"
<i>way</i>	s283	s391	any road or path affording passage from one place to another; "he said he was looking for the way out"
	s278	s386	a line leading to a place or point: "he looked the other direction"; "didn't know the way home"
<i>world</i>	s224	s696	people in general considered as a whole; "he is a hero in the eyes of the public"
	s226	s698	all of the inhabitants of the earth; "all the world loves a lover"

Figure 68 A relatively high gain in interpretational accuracy is the result of interpreting paragraphs 0 through 5 in combination.

7.6.4 Conclusions of Experiment 3

I am personally very pleased with the result of combining the first six paragraphs of Figure 52 into one. The conclusion must be that interpretational accuracy and coverage benefited a lot from this combined interpreting and it conforms very nicely with the fact that the paragraphs in combination clearly was intended as a combined whole in the textual data under the headline "Five Impulses of Civilization."

Perhaps it is not as surprising as it might appear, since indeed the small corpus obviously deals with some very restricted concepts and it makes sense that the larger the portion of text that is examined the better the result will be. Still this is only the case as long as the particular portion of text that is examined does not contain contradicting contexts or radical contextual shifts.

Most importantly, the context as represented by the interpreting graph does, I think, very well represent the context of the textual paragraphs, as it is subjectively perceived.

Find the graph and statistics of Experiment 3 in the appendices A-3 and A-5. See appendix A-4 for the glossary.

7.7 Experiment 4 - all 9 paragraphs in one sequence.

In order to further investigate whether the increase in performance of the skimming algorithm experienced in the previous section, this experiment will consider the complete corpus of Figure 52 as one paragraph. This final sequence will serve as an alternative to the structure proposed by the typography of the corpus text. The result is to be compared partly to the combined sequence of Experiment 3 and partly to the results of the isolated sequences of Experiment 1.

The entire corpus was taken from the same source namely the CivilizationIII manual and as such, the sequence cannot be regarded as the combination of completely distinct texts. Still it is interesting to see if there is enough difference in focus between the first six paragraphs in combination and the last three in isolation to cause problems for the skimming of the combined sequence.

The sequence produces only one “best” interpreting graph, just like in section 7.6. The graph scores 116 compared the sum of scores of isolated paragraphs – 53, and the sum of scores of the combined sequences 0-5 and 6, 7 and 8 in isolation – 81.

7.7.1 New lexemes in the concatenation of all sequences.

The combined sequence provokes a whole range of new concepts to be recognised. While some of these additions does make sense in the context as I perceive it from Figure 52(COMPONENT, CONQUEST are such examples) the vast majority of new lexemes are results of very dubious interpretations referring to concepts that doesn't fit very well or are decidedly misleading (the TRIAL-DEFENSE component referring to LAW SUITS is an example, but there are many almost as disruptive interpretations).

There is however one particularly interesting component in this interpretation. It involves *people, population, nation, folk, person, opponent, friend* and *world*. This component does indeed hold some important concepts to the context and is also responsible for the very respectable increase in score. Also the component containing *book-reference-text* seems pretty well motivated when reading paragraph 8, even if the interpretations offered for the words are slightly off-centre. Apart from these only the new *government-empire* component is reasonably central to the context of corpus. The *government* in the corpus does however refer to the POLITICAL IDEOLOGY OF AN GOVERNMENT, (i.e. MONARCHY, COMMUNISM, REPUBLIC, ANARCHY, DESPOTISM, DEMOCRACY etc.) rather than the RULING BODY OF A NATION like THE BUSH ADMINISTRATION for instance.

	Exp4 Ref.	
achievement	s455	the act of accomplishing something
aggression	s245	violent action that is hostile and usually unprovoked
area	s57	a subject of study; "it was his area of specialization"; "areas of interest include..."
	s59	the extent of a 2-dimensional surface enclosed within a boundary; "the area of a rectangle"; "it was about 500 square feet in area"
	s56	a particular geographical region of indefinite boundary (usually serving some special purpose or distinguished by its people or culture or geography); "it was a mountainous area"; "Bible country"
arts	s267	studies intended to provide general knowledge and intellectual skills (rather than occupational or professional skills); "the college of arts and sciences"
aspect	s330	a distinct feature or element in a problem; "he studied every facet of the question"
binding	s637	the front and back covering of a book; "the book had a leather binding"
book	s653	a number sheets (ticket or stamps etc.) bound together on one edge; "he bought a book of stamps"
	s650	the front and back covering of a book; "the book had a leather binding"
	s654	a book as a physical object: a number of pages bound together; "he used a large book as a doorstop"
	s651	a major division of a long written composition; "the book of Isaiah"
case	s72	a problem requiring investigation; "Perry Mason solved the case of the missing heir"
	s70	a person requiring professional services; "a typical case was the suburban housewife described by a marriage counselor"
	s68	a person who is subjected to experimental or other observational procedures; someone who is an object of investigation; "the subjects for this investigation were selected randomly"; "the cases that we studied were drawn from two different communities"
cash	s363	money in the form of bills or coins
challenge	s454	a demand by a sentry for a password or identification
change	s617	a thing that is different; "he inspected several changes before selecting one"
	s614	money received in return for its equivalent in a larger denomination or a different currency; "he got change for a twenty and used it to pay the taxi driver"
	s613	the balance of money received when the amount you tender is greater than the amount due; "I paid with a twenty and pocketed the change"
chapter	s594	a local branch of some fraternity or association; "he joined the Atlanta chapter"
	s596	a subdivision of a written work; usually numbered and titled; "he read a chapter every night before falling asleep"
civilization	s693	a society in an advanced state of development
component	s431	something determined in relation to something that includes it; "he wanted to feel a part of something bigger than himself"; "I read a portion of the manuscript"; "the smaller component is hard to reach"
conquest	s207	success in mastering something difficult; "the conquest of space"
	s205	an act of winning the love of someone
course	s464	a mode of action; "if you persist in that course you will surely fail"
	s458	general line of orientation: "the river takes a southern course"; "the northeastern trend of the coast"
culture	s294	the tastes in art and manners that are favored by a social group
	s292	a particular civilization at a particular stage
darkness	s40	having a dark or somber color
defense	s238	a defendant's answer or plea denying the truth of the charges against him; "he gave evidence for the defense"
demand	s156	an urgent or peremptory request; "his demands for attention were unceasing"
discovery	s198	the act of discovering something
effort	s318	a notable achievement: "the book was her finest effort"
	s316	earnest and conscientious activity intended to do or accomplish something: "made an effort to cover all the reading material"; "wished him luck in his endeavor"; "she gave it a good try"
empire	s162	a group of countries under a single authority: "the Roman empire"
feature	s52	a prominent aspect of something: "the map showed roads and other features"; "generosity is one of his best characteristics"
folk	s491	people in general; "they're just country folk"; "the common people determine the group character and preserve its customs from one generation to the next"
	s490	people descended from a common ancestor; "his family had lived in Massachusetts since the Mayflower"
force	s314	an act of aggression (as one against a person who resists); "he may accomplish by craft in the long run what he cannot do by force and violence in the short one"
	s312	physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man"
	s310	a unit that is part of some military service; "he sent Caesar a force of six thousand men"
frequency	s182	the number of occurrences within a given time period (usually 1 second); "the frequency of modulation was 40 cycles per second"
friend	s480	a person with whom you are acquainted; "I have trouble remembering the names of all my acquaintances"; "we are friends of the family"
	s478	a person you know well and regard with affection and trust; "he was my best friend at the university"
game	s684	the score needed to win a game; "he is serving for the game"
goal	s6	a successful attempt at scoring; "the winning goal came with less than a minute left to play"

	s4	the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means"
government	s679	the organization that is the governing authority of a political unit; "the government reduced taxes"; "the matter was referred to higher authorities"
greatness	s326	the property possessed by something or someone of outstanding importance
hand	s350	ability; "he wanted to try his hand at singing"
head	s520	that which is responsible for one's thoughts and feelings; the seat of the faculty of reason; "his mind wandered"; "I couldn't get his words out of my head"
	s517	the subject matter at issue; "the question of disease merits serious discussion"; "under the head of minor Roman poets"
history	s287	the discipline that records and interprets past events involving human beings: "he teaches Medieval history"; "history takes the long view"
	s286	all that is remembered of the past as preserved in writing; a body of knowledge: "the dawn of recorded history"; "from the beginning of history"
importance	s22	the quality of being important and worthy of note; "the importance of a well-balanced diet"
impulse	s19	an impelling force or strength; "the car's momentum carried it off the road"
knowledge	s362	the psychological result of perception and learning and reasoning
measure	s274	a basis for comparison; a reference point against which other things can be evaluated; "they set the measure for all subsequent work"
nation	s449	the people of a nation or country or a community of persons bound by a common heritage; "a nation of Catholics"; "the whole country worshipped him"
	s447	a federation of tribes (especially native American tribes); "the Shawnee nation"
	s448	a politically organized body of people under a single government; "the state has elected a new president"
objective	s14	the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children"
opponent	s266	someone who offers opposition
order	s299	a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today"
path	s195	a course of conduct; "the path of virtue"; "we went our separate ways"; "our paths in life led us apart"; "genius usually follows a revolutionary path"
	s194	a way especially designed for a particular use
	s193	an established line of travel or access
people	s290	members of a family line; "his people have been farmers for generations"; "are your people still alive?" (plural) any group of human beings (men or women or children) collectively; "old people"; "there were at least 200 people in the audience"
	s291	
person	s532	a human being; "there was too much for one person to do"
place	s582	an abstract mental location; "he has a special place in my thoughts"; "a place in my heart"; "a political system with no place for the less prominent groups"
	s401	a blank area; "write your name in the space provided"
	s575	the particular portion of space occupied by a physical object: "he put the lamp back in its place"
	s573	a space reserved for sitting (as in a theatre or on a train or airplane); "he booked their seats in advance"; "he sat in someone else's place"
population	s186	the people who inhabit a territory or state; "the population seemed to be well fed and clothed"
power	s145	possession of the qualities (especially mental qualities) required to do something or get something done; "danger heightened his powers of discrimination"
	s143	a state powerful enough to influence events throughout the world
problem	s559	a question raised for consideration or solution; "our homework consisted of ten problems to solve"
production	s101	the amount of an artifact that has been produced by someone or some process; "they improve their product every year"; "they export most of their agricultural production"
purpose	s109	an anticipated outcome that is intended or guides your planned actions; "his intent was to provide a new translation"; "it was created with the conscious aim of answering immediate needs"; "he made no secret of his designs"
range	s223	a variety of different things or activities; "he answered a range of questions"; "he was impressed by the range and diversity of the collection"
rate	s108	the relative speed of progress or change; "he lived at a fast pace"; "he works at a great rate"; "the pace of events accelerated"
	s107	(British) a local tax on property (usually used in the plural
	s105	a magnitude or frequency relative to a time unit; "they traveled at a rate of 55 miles per hour"; "the rate of change was faster than expected"
rear	s496	the part of something that is furthest from the normal viewer: "he stood at the back of the stage"; "it was hidden in the rear of the store"
reference	s665	a book to which you can refer for authoritative facts; "he contributed articles to the basic reference work on that topic"
	s661	an indicator that orients you generally; "it is used as a reference for comparing the heating and the electrical energy involved"
rest	s588	something left after other parts have been taken away; "there was no remainder"; "he threw away the rest"
science	s181	ability to produce solutions in some problem domain; "the skill of a well-trained boxer"; "the science of pugilism"

score	s476	a seduction culminating in sexual intercourse; "calling his seduction of the girl a 'score' was a typical example of male slang"
	s475	the act of scoring in a game or sport; "the winning score came with less than a minute left to play"
	s468	a number that expresses the accomplishment of a team or an individual in a game or contest; "the score was 7 to 0"
society	s300	an extended social group having a distinctive cultural and economic organization
	s299	a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today"
space	s402	(printing) a block of type without a raised letter; used for spacing between words
	s401	a blank area; "write your name in the space provided"
	s399	one of the areas between or below or above the lines of a musical staff; "the spaces are the notes F-A-C-E"
	s396	an area reserved for some particular purpose; "the laboratory's floor space"
speed	s114	distance travelled per unit time
	s112	a rate (usually rapid) at which something happens; "the project advanced with gratifying speed"
success	s444	an attainment that is successful; "his success in the marathon was unexpected"; "his new play was a great success"
	s340	a person with a record of successes; "his son would never be the achiever that his father was"; "only winners need apply"; "if you want to be a success you have to dress like a success"
taste	s211	delicate discrimination (especially of aesthetic values); "arrogance and lack of taste contributed to his rapid success"; "to ask at that particular time was the ultimate in bad taste"
tax	s178	charge against a citizen's person or property or activity for the support of government
territory	s44	an area of knowledge or interest; "his questions covered a lot of territory"
text	s657	a book prepared for use in schools or colleges; "his economics textbook is in its tenth edition"
	s659	the words of something written; "there were more than a thousand words of text"; "they handed out the printed text of the mayor's speech"; "he wants to reconstruct the original text"
thing	s628	an entity that is not named specifically; "I couldn't tell what the thing was"
	s624	a special objective: "the thing is to stay in bounds"
trial	s547	the act of testing something; "in the experimental trials the amount of carbon was measured separately"; "he called each flip of the coin a new trial"
	s546	the act of undergoing testing; "he survived the great test of battle"; "candidates must compete in trial of skill"
	s545	(law) legal proceedings consisting of the judicial examination of issues by a competent tribunal; "most of these complaints are settled before they go to trial"
tribe	s62	a federation (as of American Indians)
type	s565	a small block of metal bearing a raised character on one end; produces a printed character when inked and pressed on paper; "he dropped a case of type so they made him pick them up"
	s564	a subdivision of a particular kind of thing; "what type of sculpture do you prefer?"
unit	s672	an organization regarded as part of a larger social group; "the coach said the offensive unit did a good job"; "after the battle the soldier had trouble rejoining his unit"
	s669	an individual or group or structure or other entity regarded as a structural or functional constituent of a whole; "the reduced the number of units and installations"; "the word is a basic linguistic unit"
value	s273	relative darkness or lightness of a color: "I establish the colors and principal values by organizing the painting into three values--dark, medium...and light"-Joe Hing Lowe
	s272	the quality (positive or negative) that renders something desirable or valuable; "the Shakespearean Shylock is of dubious value in the modern world"
variety	s93	a category of things distinguished by some common characteristic or quality; "sculpture is a form of art"; "what kinds of desserts are there?"
	s89	a collection containing a variety of sorts of things; "a great assortment of cars was on display"; "he had a variety of disorders"
way	s195	a course of conduct; "the path of virtue"; "we went our separate ways"; "our paths in life led us apart"; "genius usually follows a revolutionary path"
	s391	any road or path affording passage from one place to another; "he said he was looking for the way out
	s386	a line leading to a place or point: "he looked the other direction"; "didn't know the way home"
winner	s340	a person with a record of successes; "his son would never be the achiever that his father was"; "only winners need apply"; "if you want to be a success you have to dress like a success"
world	s696	people in general considered as a whole; "he is a hero in the eyes of the public"
	s698	all of the inhabitants of the earth; "all the world loves a lover"

Figure 69 A relatively high number of noisy and misleading interpretations occur when the entire corpus is interpreted en bloc.

7.7.2 Lexemes that are reinterpreted to fit the context.

A few words get reinterpreted for the better as well : *course-way/path* is now very accurately interpreted with respect to the meaning that the author probably intended.

Perhaps the most remarkable result of interpreting the corpus paragraphs collectively is that *culture* is replaced from *civilization-society* by a meaning that relates to TASTE instead.

7.7.3 Conclusions of Experiment 4

When reading the corpus in Figure 52, I can readily recognize a change in focus from paragraph 5 to 6 and again in 7 and 8. As already mentioned though, they are closer related to each other than if the paragraphs had been taken from completely distinct texts and it is hard to pinpoint to exactly what should raise the alarm when studying the interpretations.

7.8 Conclusions on Experiments

I have been conducting a series of simple experiments, applying the developed prototype to the designated corpus in various set-ups. The purpose of the experiments has been to establish how the representation of contexts relates to the corpus text as perceived by a human reader. In this respect the human reader was of course myself, but I have taken great care to be as objective in my judgements as I could. Also the functioning of the chosen scoring system was to be scrutinised thoroughly in attempt to see if and how it might need modifying. A third purpose lurking in between lines has been to find out what, if any, parameters should be observed when trying to establish the turning points of contextual shifts in the data.

To take first things first, I think that the contexts represented by the various graphs in appendix A-1 generally do represent the subjective contexts (i.e.: the contexts that a human reader would most likely perceive) of the corresponding paragraphs of the corpus. There is a certain amount of noise - as could be expected, but most of the interpretations do actually have a bearing on the subjective contexts of their respective paragraphs, and a fair number of them represent concepts and concept clusters i.e.: components, that appear central to those subjective contexts. A good example is the interpretation of paragraph 0 of the corpus, commented in section 7.4.1, shown below.

One issue has drawn attention to itself throughout the experimenting, namely that two words in a given paragraph may very well be interpreted as obviously related to each other by a human reader and not by the skimming algorithm (see also section 7.4.2 and 7.4.10). This occurs when the words have meanings that are not directly related via a single relation in the MRD but instead via an implicit concept that is not explicitly realized by a word in the data. Such relationships pose a problem to the skimming algorithm since it is only allowed to check explicitly realised concepts for relations. A human reader rarely even realises the intervening concepts in these cases, but instinctively climbs the ontological edges between the concepts. I have proposed that the skimming algorithm perhaps should be allowed restricted use of such implicit concepts to produce better interpretations. While obviously facilitating more relationships to choose from, the result might

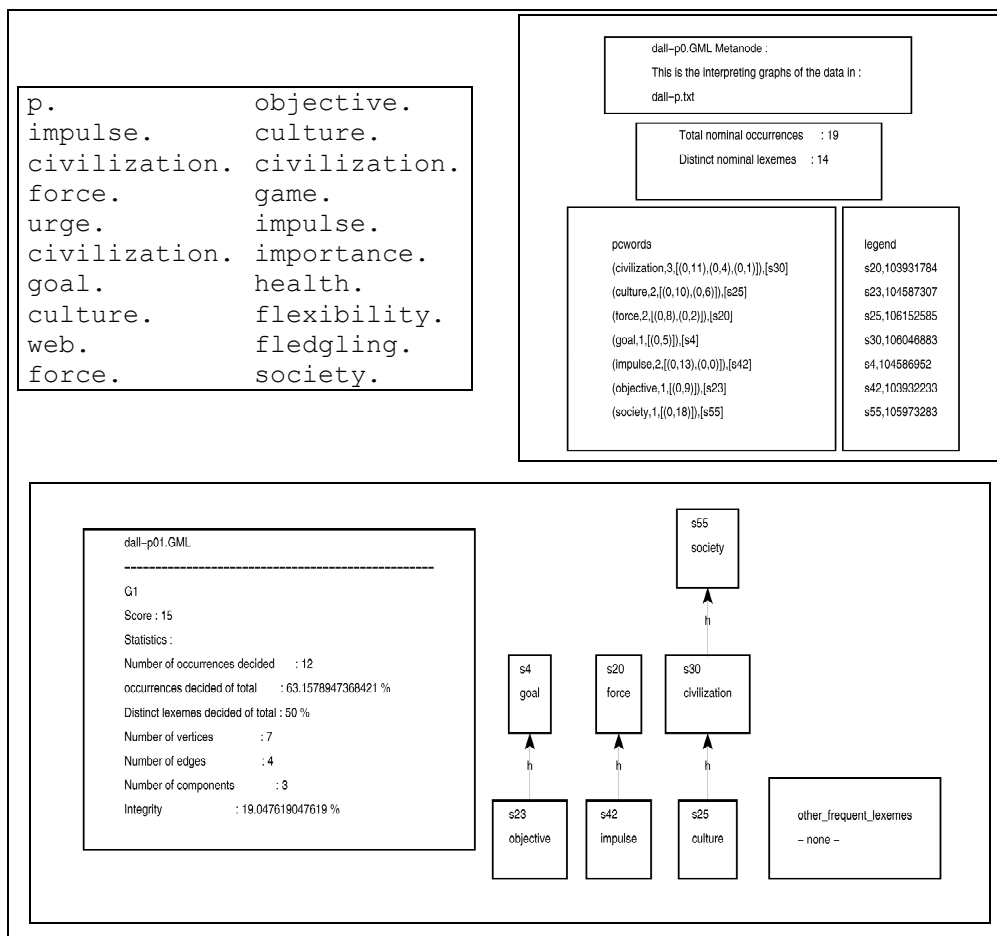


Figure 70 The graph for paragraph 0 of the corpus represents a "good" interpretation, that depicts a structure of concepts that resembles what a human reader might perceive.

clutter up from noisy relationships. The complexity of the algorithm is assumed to worsen badly (due to an increased number of edges in the CHUNK). It remains an experiment for a possible future project. The paragraph 1 of the corpus, the one named "Exploration" exemplifies a rather disappointing instance, where many "obvious" relations are missing from the interpretation shown in Figure 71 below. The glossary for both examples is present in appendix A-1

With regard to the scoring system, it seems to function almost as intended. The higher scoring graphs are generally of better quality - closer to the contexts of the paragraphs as they are subjectively perceived by at least one human reader. I have proposed that the inclusion of an edge that causes an already established lexeme to be removed from the interpretation be punished based on the importance of the lexemes involved. By importance I mean number of occurrences that will benefit from including the edge on one hand and from excluding it on the other. A modification like this will favour the first lexemes to be recognised and the most recurring ones as well. Also the positional distance between the related occurrences might be taken into consideration. The present scoring system (

- the modified scoring scheme 2) was chosen for its extreme simplicity and care has to be taken before complicating it since it does pretty well as is.

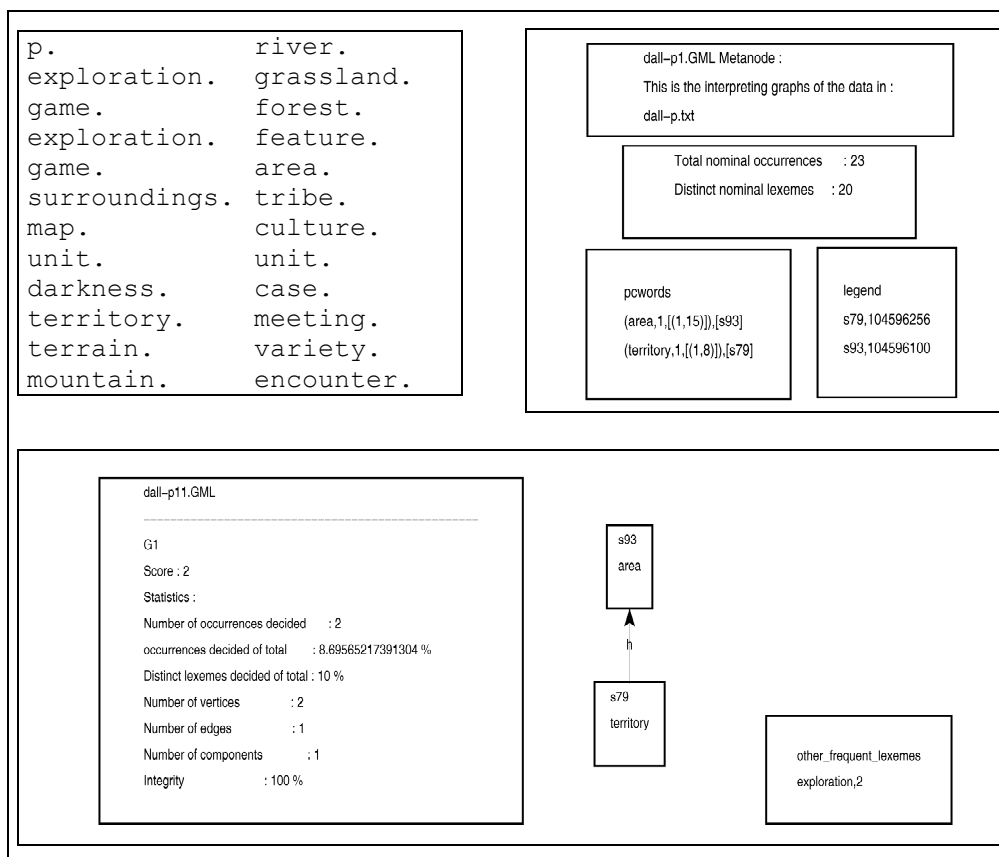


Figure 71 The graph for paragraph 1 of experiment 1, commented in section 7.4.2. *Terrain* and either of *mountain, river, grassland or forest*, are to a human reader obviously related. Never-the-less the relations are not directly present in the MRD and hence the system fails to recognise them.

Finally, I shook a stick at the problem of contextual shifts. It remains a complicated matter indeed since several scenarios can be perceived that involve contradicting interpretations. It might be a coincidental choice of words that happen to have related meanings in some obscure context far from the one at hand. It might also be the case that an implicit concept (as mentioned in section 7.4.10) present between the lines of the data just happens to be realizable by a word that occurs in the data but is intended to have another meaning. Maybe the author consciously avoided the use of the word for just that reason, or maybe indeed he included an ambiguity in his text. Allowing implicit concepts while restricting the scoring to the explicit ones could be part of a solution to this problem. Let us look a bit closer on the results of the experiments with regard to contextual shifts.

Remember that the purpose of the experiments was to see if paragraphs that fit together well semantically to a human reader, also scores

well when interpreted in combination with each other. That correlates to how well their respective contexts combines.

Sequence	Instances	lexemes	Integrity	Score	Sum of scores of isolated p's	Score pr Instance	Score pr Lexeme	Average of Sum of scores of individual paragraphs		Relative gain from joined paragraphs	
dall-p0	19	14	19%	15	-	0,79	1,07	-	-	-	-
dall-p1	23	20	100%	2	-	0,09	0,10	-	-	-	-
dall-p2	29	25	33%	5	-	0,17	0,20	-	-	-	-
dall-p3	19	18	0%	0	-	0,00	0,00	-	-	-	-
dall-p4	18	16	0%	0	-	0,00	0,00	-	-	-	-
dall-p5	17	15	19%	11	-	0,65	0,73	-	-	-	-
dall-p6	19	19	0%	0	-	0,00	0,00	-	-	-	-
dall-p7	27	26	33.3%	4	-	0,15	0,15	-	-	-	-
dall-p8	60	44	9%	16	-	0,27	0,36	-	-	-	-
								pr instance	pr lexeme	pr instance	pr lexeme
D0X-0	42	32	14%	18	17	0,43	0,56	0,40	0,53	0,02	0,03
D0X-1	48	38	11%	24	20	0,50	0,63	0,42	0,53	0,08	0,11
D0X-2	38	32	19%	15	15	0,39	0,47	0,39	0,47	0,00	0,00
D0X-3	37	28	20%	18	15	0,49	0,64	0,41	0,54	0,08	0,11
D0X-4	36	25	9%	31	26	0,86	1,24	0,72	1,04	0,14	0,20
D0X-5	35	29	19%	17	15	0,49	0,59	0,43	0,52	0,06	0,07
D0X-6	46	38	13%	21	19	0,46	0,55	0,41	0,50	0,04	0,05
D0X-7	79	56	8%	37	31	0,47	0,66	0,39	0,55	0,08	0,11
D0-5_6_7_8-0	125	93	4%	61	33	0,49	0,66	0,26	0,35	0,22	0,30
dall-0	228	160	2%	116	89	0,51	0,73	0,39	0,56	0,12	0,17

Figure 72 Master Scoring Table for all experiments and some of the statistics produced by the system. The complete table has been included in Appendix 5.

In the first part of Figure 72 the statistical results for the nine individual paragraphs of the corpus that were examined in experiment 1 are listed. The last of the **boldfaced** columns lists the individual scores of the “best” interpretation of those paragraphs respectively. These are the values that were divided by number of instances and number of lexemes to produce the score/instance and score/lexeme respectively. As such the values from experiment 1 may be accepted to represent the base results of the individual paragraphs.

The purpose of the subsequent experiments was to see if combining the individual paragraphs in various ways might lead to better scores. In this respect the combination of two paragraphs can trivially be expected to produce a score that is simply the sum of the scores of the individual paragraphs. The interesting points are those combinations that score differently than simply the sum of individual scores. To this end the average

gain in score/instance and score/lexeme computed, listed in the final column of Figure 72.

A positive value here indicates that the scoring system preferred the particular paragraphs in combination rather than in isolation. The bigger the value the higher the preference.

A 0.00 means that the scoring system has no preference with respect to interpreting the paragraphs in combination or isolation.

A negative value would mean that the paragraphs were better off being interpreted in isolation.

The layout of the corpus, as presented in Figure 52, suggests that paragraphs 0 should connect better with paragraphs 1-5 than with paragraphs 6,7 or 8.

Experiment 2 was designed to see if such a connection could be recognised by the scoring system. 8 sequences was constructed from combining paragraph 0 with each of paragraphs 1 through 8.

In Figure 72 the scores of interpretations of these sequences are listed as d0X-0 ... d0X-7. If the first 5 of the scored significantly better on the average than the last 3 it would be an indication that the system might be able recognize the described connection between the paragraphs of the corpus.

Regarding the values in Figure 72, no such distinction can be made on the present grounds. It is worth to mention, however, that none of the combinations contained significant contradictions either. This maybe reflects the fact that after all, the paragraphs were taken from the same portion of the same text.

For the final experiment I constructed a new sequence from combining all paragraphs 0 through 5 into one, d0-5_6_7_8-0 in Figure 72. Furthermore I constructed a new sequence from combining all 9 paragraphs into one big sequence, dall-0 in Figure 72.

It is clear to a human reader that the first 6 paragraphs of the corpus does form a connected whole distinct from the last three. The experiment is to see if this is reflected in the way the skimmer performs on these two sequences.

The average gain in score/lexeme in the interpretation of the combination of paragraphs 0-5 is 0.30. This is much better than any of the sequences of experiment 2, and also significantly better than the 0.17 from the interpretation of all 9 paragraphs in combination.

This to me clearly suggest that the system does indeed prefer the semantic line to be drawn between paragraphs 5 and 6 just as the layout of the corpus indicates to a human reader.

While none of this represents conclusive fact or proof in any way, it does indicate that further experiments with - and development of the idea of skimming might very well prove fruitful:

Experiments that compare a large number texts, both of individually distinct and similar concepts, could possibly confirm or refine my findings.

Experiments involving referees that are not as close to the development of the project as myself would also be desirable. Even multiple referees that were to decide subjectively on the context of the textual data. Their findings should then be compared to the results of the algorithmically found contexts.

Finally experiments involving larger portions of text on a restricted context should be conducted. Experiment 3 clearly indicates that relatively long expositions within a given context yield better results when skimmed than relatively shorter ones. None of the paragraphs in the chosen corpus are very long - to say the least - and the context is shifting slightly but rapidly throughout the thing.

All in all I think the prototype performed as well as could be expected on the data available.

8 Concluding remarks and further development

This project draws to an end and it is time to make out the state of things. The present paper is intended to represent an attempt to address a small but important corner of the problem of automatic natural language understanding. Apart from the core issues of logic, lexical semantics and pragmatics, the ideas developed in the thesis are of course my own. They were gradually developed through my formal training at DIKU and IAAS. While working on the project, it has however, come to my knowledge that other researchers have recently begun investigating very similar ideas. While this of course encouraged me to press on along my own lines, it also presented a pressing need for making it completely clear what my contribution involves and indeed to describe very carefully what inspired these ideas. As such, I think that the project presented a unique opportunity for me to try my hand at the complex discipline of research. I have from the very start made it a goal in itself to hone my researching skills and present as sound work as possible. In this respect, I have taken great care to make proper references whenever I was consciously inspired by the work of other people, and it is my sincere hope that I did succeed in doing so.

To the very best of my knowledge, this project is the only one that actually implements and documents a working experimental prototype that addresses lexical disambiguation, context recognition and interpretation through lexical relationships and pragmatics.

8.1 Related research

As already mentioned several projects drew attention to themselves as I researched for my own project. In this section of my thesis I will present two of the most important ones.

8.1.1 The Generative Lexicon)

In the words of James Pustejovsky himself “The Generative Lexicon”, (Pustejovsky, James 1995), is a study of

... the interaction of word meaning and compositionality as they relate to the issues of

- *the creative use of words in novel contexts;*
- *an evaluation of lexical semantic models on the basis of compositionality.*

The book represents a very thorough study of a broad selection of lexical semantic aspects with many intriguing views and it leads to far to touch on everything treated in the book. Especially one main argument of the book does however relate very closely to the ideas described in my own work. It concerns how the lexicon is organised to represent polysemy, i.e.: different meanings of the same word. He argues that

Computational and theoretical linguists have largely treated the lexicon as a static set of word senses, tagged with features for syntactic, morphological and semantic and semantic information.

In order to facilitate dynamicity on the part of the lexicon so it can properly accept new meanings and new words as well, he proposes a very elaborate cluster of structures for lexical entries including:

- *Argument Structure (for the representation of adicity information for functional elements),*
- *Event Structure (for the representation of information related to Aktionsarten and event type, in the sense of (Vendler, Z. 1967), and related work),*
- *Qualia Structure (for the representation of the defining attributes of an object such as its constituents parts, purpose and function, mode of creation, etc.),*
- *Inheritance Structure (for the representation of the relation between the lexical item and others in the lexicon).*

In essence he argues that to properly realize the subtle differences between concepts they need to be thoroughly described in the lexicon. Of the structures just referred to, the *Qualia Structure* is especially appealing with regard to my own ideas. Pustejovsky describes it as specifying four essential aspects (also called *quale's*) of a word's meaning (i.e.: the concept realized by the word, in my terms):

- *Constitutive aspect: the relation between an object and its constitutive parts;*
- *Formal aspect: that which distinguishes it within a larger domain;*
- *Telic aspect: its purpose and function;*
- *Agentive aspect: factors involved in its origin or "bringing about".*

He makes a point of noting that every category (word class) expresses a Qualia Structure, while not all lexical items carry a value for each qualia aspect. Each syntactic category is studied in these terms.

As a part of a much larger goal, Pustejovsky suggests that in addition to the traditional relationships of hyponymy, synonymy, antonymy and meronymy, nominal concepts should also be associated with relations to purpose and function, what process created it of what materials and by what or whom. So the word *book* should in one of its interpretations relate to the *physical realization* of the book itself, that it *holds information*, that it was created in a *writing-process* in order for people to *read* it.

The qualia examples below were all taken from Pustejovsky's book, (Pustejovsky, James 1995). In the examples the constitutive quale is not included. The constitutive quale, in Pustejovsky's terms, express the kind

of information that WordNet record as meronymy. Ideally each entry of a lexical item should include a constitutive quale for each dot-member of a complex type (see below). Having mentioned this, I will present the examples as they appear in Pustejovsky's book, i.e.: without the constitutive quale.

```
[book
  ARG STRUCT = [
    ARG1 = x:info
    ARG2 = y:physobj
  ]
  QUALIA = [
    info.physobj_lcp
    FORMAL = hold(x,y)
    TELIC = read(e,w,x,y)
    AGENT = write(e',v,x,y)
  ]
]
```

Figure 73 Part of the lexical entry for the word *book* as suggested by Pustejovsky. There should have been a constitutive quale here stating that the *physobj* consists of pages of paper and a binding with a front and back, etc. Whereas the *info* dot-member is made up of chapters, paragraphs and so on.

From the structure above it can be seen that Pustejovsky's idea involves a very rich type-system. The *lexical conceptualisation paradigm* (lcp) for the word *book* has a complex type referring to both a type called *info*, the one called *physobj* and finally also as the combination of the two referred to as *info.physobj*, i.e.: with the "dot" resembling the symbol for functional composition (•). This was done to explain how both facets are available for interpretation in the example :

"Jenny liked the book"

Here Jenny might both like the contents of the book, i.e.: the information that it contains, or she might like physical book itself, for its fine binding or because it was given to her by her mother or for some other reason. Finally Jenny might not intend to distinguish either of the two facets in particular. All three readings of the word is available at the same time hence the complex type. Such complex types are represented by so-called DOT-objects or dotted types like the *info.physobj* in this example. The relation between the two *dot-members* (aspects) of the word is also present, namely that the physical book (*y*) holds the information (*x*). The purpose of the book is recorded as the telic *quale* namely a reading event (*e*) of some person (*w*) reading the book (*x.y*). Finally the constitutive origin of the book is recorded in the agentive *quale*, a writing event (*e'*) where some person (*v*) wrote the book (*x.y*).

The type system involves that one type may be a subtype of another type. It is fair to say that music is a kind of information, for instance. This way the type system itself expresses some of the relational information between the various lcp's and indeed the theory involves a set of relations over types. I will not go into detail on this matter here, but think it should be mentioned.

As a contrastive example of qualia, Pustejovsky suggests this lexical entry for the word *symphony*.

```
[symphony
  ARG STRUC =      [
                    ARG1 = x:music
                    ]
  EVENT STRUCT =   [
                    E1 = e1:process
                    ]
  QUALIA =          [
                    music.process_lcp
                    FORMAL = perform(e1,w,x)
                    TELIC = listen(e',z,e1)
                    AGENT = compose(e'',y,x)
                    ]
]
```

Figure 74 The lexical entry for the word *symphony* as suggested by Pustejovsky. Again the constitutive quale is missing from this example. Also it seems that an alternative ARG2 is missing, referring to the physobj clearly available for this word apart from the music.

Here the Qualia makes it possible to distinguish between the composing (e''), the listening (e') and the performing (e₁) events of the symphony. Note that the music_lcp is a subtype of the info_lcp, and that the perform-event has type process_lcp.

As a final example of the Qualia information proposed as a part of the lexicon I will present the lexical entry for the word *newspaper*. The notion of newspaper seems very close to that of book. But consider the below example taken from Pustejovsky.

- a) Jenny sued the newspaper.
- b) ! Jenny sued the book .

This example illustrates that there is a reading of newspaper that represents the organisation publishing the newspaper. This reading is however not available to book. Pustejovsky suggest that newspaper be represented in the lexicon as follows.

```
[newspaper
  ARG STRUCT =      [
                    ARG1 = x:org
                    ARG2 = y:info.physobj
                    ]
  QUALIA =          [
                    org.info.physobj_lcp
                    FORMAL = y
                    TELIC = read(e2,w,y)
                    AGENT = publish(e1,x,y)
                    ]
]
```

Figure 75 The lexical entry for the word *newspaper* according to Pustejovsky. This example is also missing the constitutive quale.

So when the “newspaper is sued”, the verb, *sue*, selects for an argument of an appropriate type, i.e.: one that can be sued. Here x:org has a suitable type for suing. The “extra” reading of newspaper is represented by the third dot-member of type org_lcp

I can not possibly do justice to a book in excess of 200 packed pages. I would, however, like to point out that Pustejovsky proposes a very rich variety of information to be available in the lexicon (or at least in close relation to it, see below). With this information he achieves some very interesting theoretical results.

With regard to my project it is certain that if indeed the MRD included - or held reference to - all the information that Pustejovsky suggests, the skimmer would quite likely benefit a lot with respect to precision and degree of interpretation.

I must also point out that later work suggests that the information represented by the Qualia be removed from the lexicon itself, while retaining a close relationship to it. One of the reasons why this may be so is a problem distinguishing between the alternative dot-members of a complex type. See for instance (Asher and Pustejovsky 2000), that also addresses the point that usually a book bought in a bookstore isn't really written explicitly, but rather printed and published – while surely the author wrote it.

It leads to far to pursue these problems in this work. I will conclude my reference to the Generative Lexicon by emphasizing that to my mind, Pustejovsky's book, (Pustejovsky, James 1995), remain an extremely interesting exposition on the nature of the meaning of words and the organisation of the lexicon.

8.1.2 Ontoquery

The other recent project that relates to the ideas studied in the present paper is the OntoQuery research project. The project represents the combined effort of the five partners of the project:

- Roskilde University (www.ruc.dk)
- Technical University of Denmark (www.dtu.dk)
- Copenhagen Business School (www.cbs.dk)
- Centre of Language Technology (www.cst.dk)
- University of Southern Denmark (www.sdu.dk)

The following excerpt is from “Presentation of the OntoQuery project” Erdman Thomsen, (Erdman Thomsen, Hanne et al 2001):

OntoQuery addresses content-based retrieval processes and access to Danish text sources such as online document databases and encyclopaedias, making use of limited computational natural language understanding. The overall goal of the project is to develop a general theory for:

- *ontological representation of domain knowledge,*
- *ontological semantics for natural language phrases, and*
- *ontology-based search in text databases.*

The project developed a query language named Ontolog for searching in text documents for its content. The main goal remained to facilitate search and retrieval of relevant texts from a database, focusing on semantic, ontological qualities of the text compared to the sought after qualities expressed in the query. In this respect, relations between concepts, like the ones I am studying in my paper, also plays a central role for OntoQuery.

The wide variety of relations between concepts suggested by OntoQuery compares nicely to the ones suggested by Pustejovsky.

While the OntoQuery project differs somewhat from the skimming project, their common denominator, the focus on conceptual relations and organisation, clearly occupies both.

Again, it is impossible to do justice to the project in just a few paragraphs and the reader is encouraged to look for more information on the official homepage (www.ontoquery.dk). The OntoQuery project has had considerable success in achieving its goals and several prototypes has been implemented. See also (Jensen and Skadhauge 2001), (Erdman Thomsen et al 2001).

8.2 Evaluation of the project with regard to original goals

Originally I set out on the theory that the *perceived subjective context* of a portion of natural language text closely relates to a reasonably restricted *set of core concepts and particular words realizing them*. (see appendix A-0 for the original project description).

I intended to investigate what was needed to design and implement an algorithm, that was able to accept natural language text as an argument and produce a recognizable context for that data, represented by a set of semantically connected lexemes. The rationale of this was inspired by the *cooperative principle* of Grice indicating, that the most likely context of the text would probably involve those interpretations of the words that were related closer to each other semantically.

Drawing upon this experimental prototype I wanted to gain information as to what behaviour in the textual data might signal contextual shifts and if such behaviour could be formalised.

As the work progressed several important issues was encountered and addressed. I will summarise the most important ones here.

8.2.1 Sequences of nouns instead of full text.

Early on I realised that if I was to get to the semantic reasoning that I found to be at the centre of the problem, I would have to restrict myself. Most importantly I didn't want to spend energy on solving problems that has already been solved satisfactory. I particular problems like syntactical

and morphological analysis seemed to be outside the scope of the formal semantic problem that was my concern. To avoid dealing with these problems I restricted the scope of the thesis to concern the nouns of the textual data only and employed several third party tools to transform the text into a set of nouns in their orthographical form.

The TOSCA-ICLE tagger was described and applied to a sample portion of the chosen corpus. I presented arguments that a fairly simple rule-governed algorithm could be designed that transformed the tagger output into the desired noun sets. I did however not implement such an algorithm, because its functioning could easily be mimicked by hand to suit my experimental purposes. If the algorithm is to be applied to large corpora actual implementation of the process is of course necessary. For the small corpus presented and experimented with here doing the transformation by hand posed no problems but the one of avoiding typos.

8.2.2 Restricted lexical relationships.

Still trying to keep the scope of the project as focused as possible, I employed the WordNet dictionary to serve as a reference for word meanings and lexical relationships. WordNet offers a variety of lexical relationships. I chose to regard most of those relationships that pertain to nouns, namely hyponymy, meronymy and antonymy. Later on I added synonymy as well. These relationships relate to the ontological structure of concepts represented by noun meanings. In order to keep things as simple as possible I decided to restrict the algorithm to take only one-step relations into consideration. In theory all concepts are related to each other in that they are all CONCEPTS so to speak. I wanted to consider relationships between concepts realised by words in the data only and with no intervening relationships.

This restriction did have some consequence to the interpretations of the algorithm. Some obviously related words were not recognised as such most likely because of intervening relationships involving concepts that were not realised in the text. See Figure 76 for an example.

An obvious and necessary expansion of the algorithm involves taking more remote relations into consideration as well and to figure out just how far out the relationship can be while still subjectively perceived.

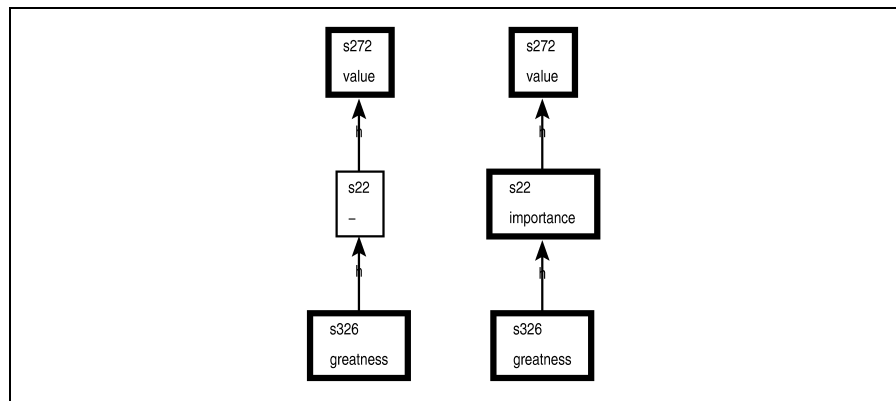


Figure 76 An illustration of the notion of implicit intervening concepts. In paragraph 5 is mention of the words *value* and *greatness* (left graph). The words have related meanings namely s272 and s326 via s22. Only s22 has no realizations in paragraph 5. I say that s22 is an implicit concept of the paragraph if a human reader would the relationship as part of the context of the paragraph. The relationship is not recognised by the skimmer until paragraphs 0 and 5 are interpreted in combination (right graph). Paragraph 0 contains the word *importance* that realizes the concept. The legend used in this figure is that of experiment 4.

Finally WordNet does not contain relationships that relate objects to their standard use, for example HAMMER to CARPENTRY, or LURE to FISHING. Neither is there mention of the relationship between BOOK and AUTHOR or WRITING. These relationships and more like them are however extremely important in order to make the proper decisions as to what a text is about. James Pustejovsky concerns himself intensely with these in his book “The Generative Lexicon”, as already mentioned. I personally agree that NLP algorithms can never be better than the lexicon they employ, and indeed that that the more relational information available via the lexicon, the greater the usability.

8.2.3 The scoring scheme.

In order to facilitate the comparison of interpretations necessary to decide on the “better” of several alternatives I introduced a very simple scoring system. Each occurrence of a sequence included in an interpretation would add 1 to the score of that interpretation every time the respective concept participated in an relationship in the interpretation. I argued that this scoring scheme would behave almost as if all the cooperative indicators inspired from Grice had been taken into consideration. While I have not proposed proof of this I have supplied circumstantial evidence that the simple scoring scheme does pretty well choose the better interpretations. (There were horrible exceptions, however). Future work must include experimenting with alternative scoring schemes, but for the purpose of the experimental prototype developed in this project, the simple scoring scheme proved sufficient.

8.2.4 Tracking contextual shifts

The prototype implementation of the skimmer never reached the degree of sophistication needed for dynamic comparison between the *context already identified* and the *context currently under consideration*. It was my aim to examine what would be needed in order to decide whether the two were instances of the same or if the context had shifted.

Indeed several things may be the case when a lexeme doesn't seem to fit with the established context of a text. In Figure 75 below the top left square represents the state of affairs where words have their predicted meanings that pose no surprises with respect to the established and now familiar context. When a word cannot be interpreted to fit the familiar context, it might be a signal of at least two different events, represented by the transitions out of the top left area. Either the context has shifted and subsequent words should be interpreted accordingly, or the author invented a new meaning for a word that did not until now have make sense in the perceived context. This should ideally result in an update in the MRD and the context as well. The experimental prototype was never intended to decide on what kind of transition is at play, let alone suggest actions (though, that of course is among the essential abilities of a complete natural language understanding system, but that is still utopia). Instead I looked for indications that something out of the ordinary might be the case when transitions occurs. The bottom right area of Figure 77 pertains to creative word use in novel contexts. This is one of the focal points of Pustejovsky, that also inspired this figure. (The skimming project was never intended to explore that wilderness at all.)

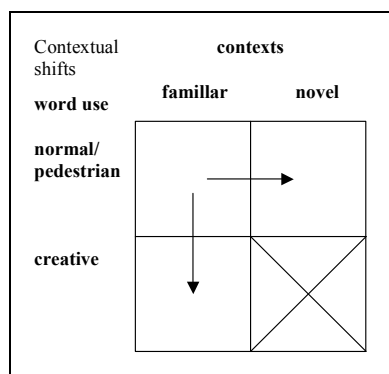


Figure 77: An abstraction of different words uses with respect to different contexts.

In order to perform the thought experiment, I compared interpretations of the corpus in several compositions and sizes. The result vaguely indicated that indeed the average score per lexeme in the four experiments did hold some of the desired information that might represent a signal of a possible transition. Especially when comparing the context of paragraphs 1-5 as a whole to the contexts of the all paragraphs in isolation.

While this compared surprisingly nice with the perceived structure of the small corpus, it is apparent that the corpus chosen for this project is far

to small and involve contexts far to similar to facilitate any conclusive results. Future research should address this by comparing significantly larger and contextually more diverse textual excerpts.

8.3 Conclusion

The real purpose of the thesis has been to experiment with the idea that lexical relationships, word use and semantic context influence each other. I wanted to see if this fuzzy intuition could be formalised to a degree where it could be implemented and applied to actual occurring linguistic data.

Even though I performed quite a few shortcuts along the way, skipping morphology and syntax altogether, the prototype program, based on the CHUNK-representation, the experimental algorithm and scoring-scheme does represent an faithful implementation of the original idea.

The question that remains to be answered is : is it all really good for anything?

On the downside, it seems clear that in its current state the prototype program is not very useful for at least the following reasons:

- Only simple relations of limited coverage are analysed.
- Only explicitly realized concepts are analysed.
- Necessary dynamic behaviour remains to be implemented.

It is clear that a useful implementation of the skimmer must allow for concepts to be related transitively.

Given for instance concepts a, b and c and some relation R, we should have that:

$$R(a,b) \wedge R(b,c) \rightarrow R(a,c)$$

even if only concepts a and c are actually realised in the text. The human mind easily and readily recognises *complex* relationships between concepts - composed of several distinct relations. The system as is, only recognises such relations IF all the involved concepts are explicitly realised in the data and this is too restricted taking the granularity of the MRD into consideration as well. An example of relationships that should rightly be considered by the system, is the blatantly unrecognised relationships between certain interpretations of *terrain* and for instance *forest* or *grassland* in the second paragraph of the corpus. (section 7.4.2 and also Figure A-1.2 of the appendices.)

I have already mentioned how the system would benefit from access to information like the kind contained in the Qualia as suggested by Pustejovsky.

On the other hand the system was not intended to represent a finished polished product per se, but rather a way of experimenting with central issues of lexical semantics and how semantic contexts might be realised and accessed. As such I think the prototype succeeds to significant degree:

- The prototype does automatically produce (sparse) semantic networks as possible interpretations of arbitrary textual data. Furthermore the emerging semantic networks could be stored for future matching and/or elaboration, almost as a memory of experienced domains. I.e.: “does this text look like anything I’ve ever seen ?”)
- The scoring system does represent at least a sketchy suggestion of how pragmatic measures may be incorporated in the evaluation and comparison of interpretations.
- Even though there was a large degree of “noise” many important concepts were recognised for the right reasons i.e.: their central place and relations in the particular semantic contexts.
- The present work is a sketch that allows and invites for expansion along several different dimensions.

The current system accepts sequences of various length, separated by the “p.”- marker and interprets them en bloc. As such the results of the experiments represent grounds for experimenting with contextual boundaries in textual data. Just when can one expect two sequences to treat the same context - and when is it likely that they treat different contexts? This might be addressed by expanding the prototype to work dynamically one word at a time, instead of a sequence at a time. A dynamic skimmer might be set to “decide” - word for word - whether it should be interpreted in conjunction with the “current” context or if it should form its own “new” context.

It is clear that the work has just begun and that the project does not present any conclusive proofs.

I would like to see experiments with the prototype on much larger and more semantically diverse corpora to support the current findings. I would also like the prototype to be expanded to take into regard additional relationships like the ones proposed by Pustejovsky and also allow for complex transitive relationships between concepts. Finally I think that a dynamic version of the prototype would prove extremely useful from a theoretic point of view. However these issues will have to be addressed in later projects as this one has already greatly outgrown its shoes, so to speak.

Never the less, I am satisfied that I did put a (tiny) scratch in the marble and it seems clear to me that the developed prototype does touch upon very central problems of natural language understanding. Problems that any similar application must address at the very least. I find that there are reasonably clear indications of the theoretical soundness of employing

lexical relations in the interpretation, disambiguation and context recognition of natural language.

As I researched the ideas involved in this paper it became apparent that there is currently a rising interest in natural language understanding and computational linguistics throughout the fields of computer science, cognitive science and logic. This hopefully bodes well for the future progress in this area of research.

References

- Allen, James (1994)
Natural Language Understanding
Benjamin/Cummings Publishing Company Inc.
- Allerton, D. J. et al (1979)
Function and Context in Linguistic Analysis
Cambridge University Press.
- Asher, Nicholas and Pustejovsky, James (2000)
The Metaphysics of Words in Context
Mr. Asher's homepage:
www.utexas.edu/cola/depts/philosophy/faculty/asher/papers.htm .
- Bratko, Ivan (1990)
PROLOG - Programming for artificial intelligence, 2nd ed
Addison-Wesley Publishes Ltd.
- Cruse, D. A. (1986)
Lexical Semantics
Cambridge University Press.
- Dalrymple, M., Kaplan, R. M., Maxwell III, J. T. and Zaene, A. (1995)
Formal Issues in Lexical-functional Grammar
CSLI Lecture Notes No. 47 .
- Erdman Thomsen, Hanne et al (2001)
Ontologies and Search.
Proceedings of the 2nd OntoQuery Workshop - Lamda No. 28,
Copenhagen Business School .
- Grice, H. P. (1975)
Logic and Conversation
In Cole Morgan 1975 pp.41-58.
- Haan, P. de and Halteren, H. van 1997
The TOSCA-ICLE Tagset
University of Nijmegen.
- Haas, W. (1962)
The Theory of Translation
Oxford University Press .
- Infogames Interactive (publ) (2001)
Civilization III - Instruction Manual .
- Jensen, Per A. and Skadhauge, Peter (editors) (2001)
Ontology-based Interpretation of Noun Phrases.
Proceedings of the first International OntoQuery Workshop
University of Southern Denmark .
- King, Stephen (1987)
The Eyes of the Dragon
Nal Penguin Inc.
- Loukides, M. and Estabrook, G. (editors) (1999)
Unix in a Nutshell
O'Reilly & Associates, Inc.

- Nilsson, Nils J. (1998)
 Artificial Intelligence - a new synthesis
 Morgan Kaufmann Publishers, Inc.
- O' Grady, W., Dobrovsky, M. and Katamba, F. (1997)
 Contemporary Linguistics : an introduction
 Addison Wesley Longman Ltd.
- Pustejovsky, James (1995)
 The Generative Lexicon
 MIT Press.
- Saeed, John. I. (1997)
 Semantics
 Blackwell Publishers Ltd.
- Sag, Ivan. A. and Wasow, Thomas (1999)
 Syntactic Theory, a formal Introduction
 CSLI Lecture Notes No. 92 .
- Saussure, Ferdiand de (1915)
 Cours de Linguistique Générale. Pyot. Paris.
 In translation:
 Course in general Linguistics, 1974, Fontana/Collins .
- Swart, Henriette de (1998)
 Introduction to natural language semantics
 CSLI Lecture Notes No. 80 .
- Vendler, Z. (1967)
 Linguistics in Philosophy Cornell University Press
- Verschueren, Jeff (1999)
 Understanding Pragmatics Arnold Publishers.

Appendices

A-0 Project description

Arbejdsbeskrivelse for tværfagligt speciale inddragende datalogiske og lingvistiske problemstillinger og metoder. Specialet skrives på engelsk.

Udnyttelse af semantiske relationer mellem nominer til registrering af topic-change¹¹ i løbende natursprogstekst som redskab til identifikation af kontekst.

Identifying context through tracking of topic-change in running natural language text using nominal semantic relationships.

Af Stud. Scient. Ole Torp Lassen, DIKU forår 2004

: _____

Konsulent : Neil D. Jones¹²

: _____

Hvorfor :

Udvikling af metoder og systemer til registrering af betydning i natursproglige korpora har længe udfordret fagkundskaben indenfor både datalogi, sprogvidenskab samt kognitionsvidenskab. Efterhånden er en god forståelse opnået for problemstillingerne forbundet med denne opgave.

Samtidig er nødvendigheden og efterspørgslen på effektive systemer til sprogbehandling blevet mere og mere udtalt. Frasortering a e-mail spam, artikel-søgning, automatisk informations-udledning, maskinoversættelse, automatisk resumé- og indeks-generering er vel oplagte eksempler på anvendelsesområder, der stiller krav om metoder til effektiv natursprogsforståelse.

Udviklingen af effektive systemer til natursprogsforståelse og tekstkategorisering bremses imidlertid noget af tilsyneladende uomgængelige høje kompleksitets-forhold , bl.a. i forbindelse med afgørelse af flertydigheder og kreativ sprogbrug .

Det er specialets tese at der er sammenhæng mellem nominers semantiske internrelationer og den ontologiske kontekst, som de indgår i og at denne sammenhæng kan udnyttes til begrænsning af NLP-systemers kompleksitet. Det er specialets mål at undersøge i hvor vid udtrækning tesen kan bekræftes, og desuden at inddrage metoder og teorier fra både datalogi og sprogvidenskab i et frugtbart samspil.

Hvad :

Nye forskningsteorier og –resultater fra datalingvistik, kognitionsvidenskab og sprogfilosofi peger på sprogforståelses-paradigmer der hviler på ontologibaseret begrebsanalyse som en mulig vej ud af vildniset. Jeg ønsker at analysere og eksperimentere med udviklingen af et system til automatisk genkendelse af den ontologiske kontekst for en vilkårlig formelt informerende natursprogstekst.

Målsætning :

Jeg forestiller mig udviklingen af en række trinvist forfinede prototyper til automatisk registrering af topic-change¹ i faktisk forekommende data fra natursproglige korpora. Som sidste prototype tænker jeg mig et system, der som inddata tager uddrag fra manualen for pc-spillet Civilization III. Som uddata for systemet forventer jeg en inddeling af de i uddraget indgående nominer i sammenhængende grupper, der kan fortolkes som svar på spørgsmålet:

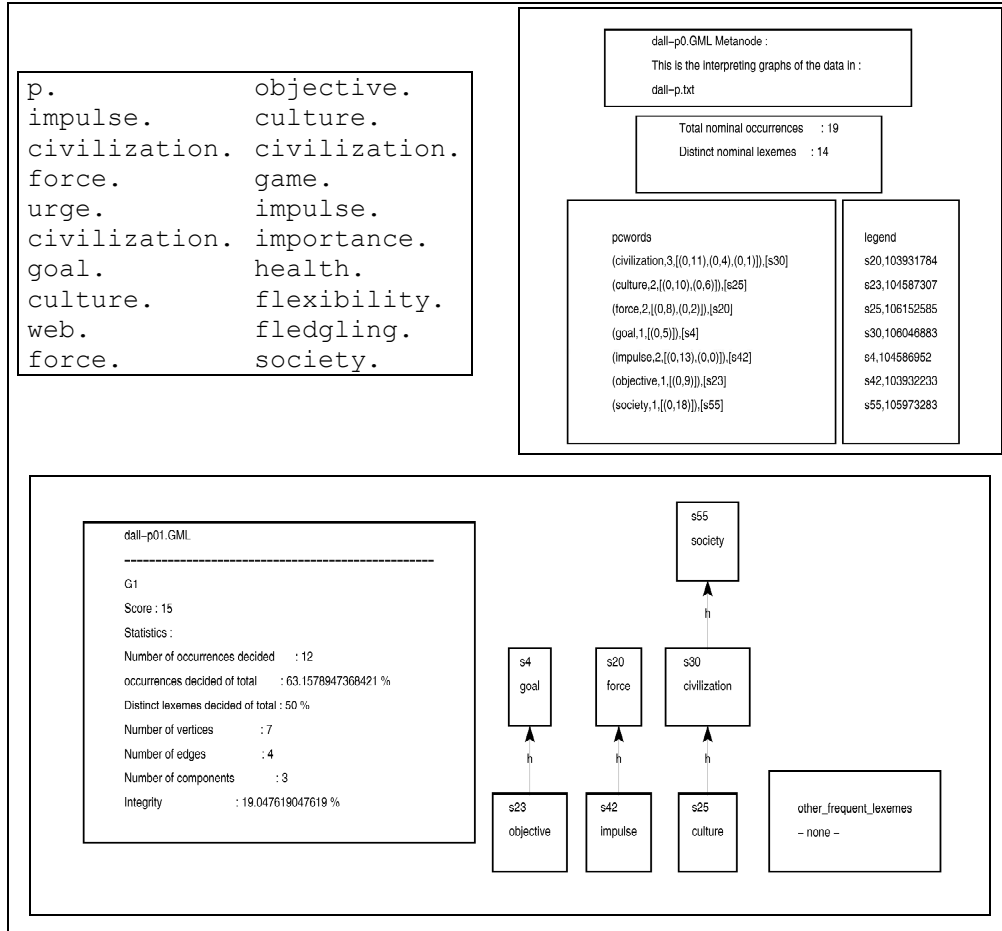
”Hvad handler uddraget om ?”

Specialet fokuserer på de semantiske relationer mellem nominerne i behandlingen af inddata, snarere end på statistiske sammenhænge mellem ordbrug og tolkning. Rimeligt succesrige systemer beroende hovedsagligt på statistiske metoder, er udviklet. Det er imidlertid forfatterens mening at de to angrebsvinkler ville drage stor nytte af et samspil med hinanden, og at ingen af dem kan lades ude af betragtning for et fuldstændigt automatisk sprogforståelsessystem. Dette motiverer en undersøgelse af betydningen af semantiske relationer for begrebsanalyse og kontekstidentifikation.

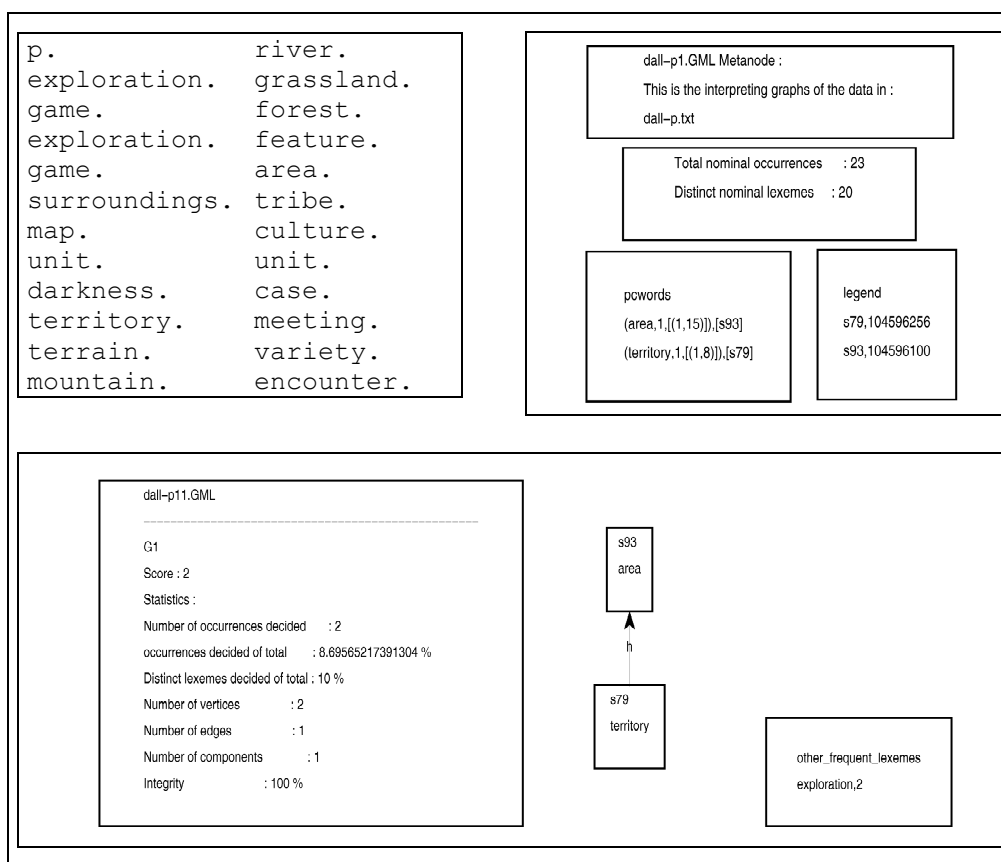
¹¹ Begrebet *topic-change* savner mundret dansk ækvivalent. Nærmeste danske betegnelse er emneskift eller fokusskift.

¹² Også adjunkt Peter Rossen Skadhauge fra Institut for datalingvistik ved Handelshøjskolen i København blev tilknyttet projektet.

A-1 Experiment 1 graphs



A-1.1: Sequence pertaining to the first paragraph of the textual data and its corresponding GML-files.



A-1.2 The second paragraph. (game should also have been mentioned among frequent lexemes)

p. tilt.
civilization. economy.
complexity. cash.
production. cow.
resource. happiness.
tax. population.
rate. population.
terrain. entertainment.
purpose. unrest.
speed. presence.
population. power.
city. luxury.
goods. resource.
tax. demand.
science. empire.

dall-p2.GML Metanode :
This is the interpreting graphs of the data in :
dall-p.txt

Total nominal occurrences : 29
Distinct nominal lexemes : 25

pcowords
(power,1,[(2,24)],s211)
(rate,1,[(2,5)],s157,s156,s154)
(science,1,[(2,13)],s174)
(speed,1,[(2,8)],s164,s162)
(tax,2,[(2,12),(2,4)],s171)

legend
s154,110990504
s156,109585403
s157,103948875
s162,110978183
s164,103948579
s171,109580808
s174,104350965
s211,104349777

dall-p21.GML

G1
Score : 5
Statistics :
Number of occurrences decided : 5
occurrences decided of total : 17.2413793103448 %
Distinct lexemes decided of total : 16 %
Number of vertices : 4
Number of edges : 2
Number of components : 2
Integrity : 33.3333333333333 %

dall-p22.GML

G2
Score : 4
Statistics :
Number of occurrences decided : 4
occurrences decided of total : 13.7931034482759 %
Distinct lexemes decided of total : 16 %
Number of vertices : 4
Number of edges : 2
Number of components : 2
Integrity : 33.3333333333333 %

dall-p23.GML

G3
Score : 4
Statistics :
Number of occurrences decided : 4
occurrences decided of total : 13.7931034482759 %
Distinct lexemes decided of total : 16 %
Number of vertices : 4
Number of edges : 2
Number of components : 2
Integrity : 33.3333333333333 %

A-1.3

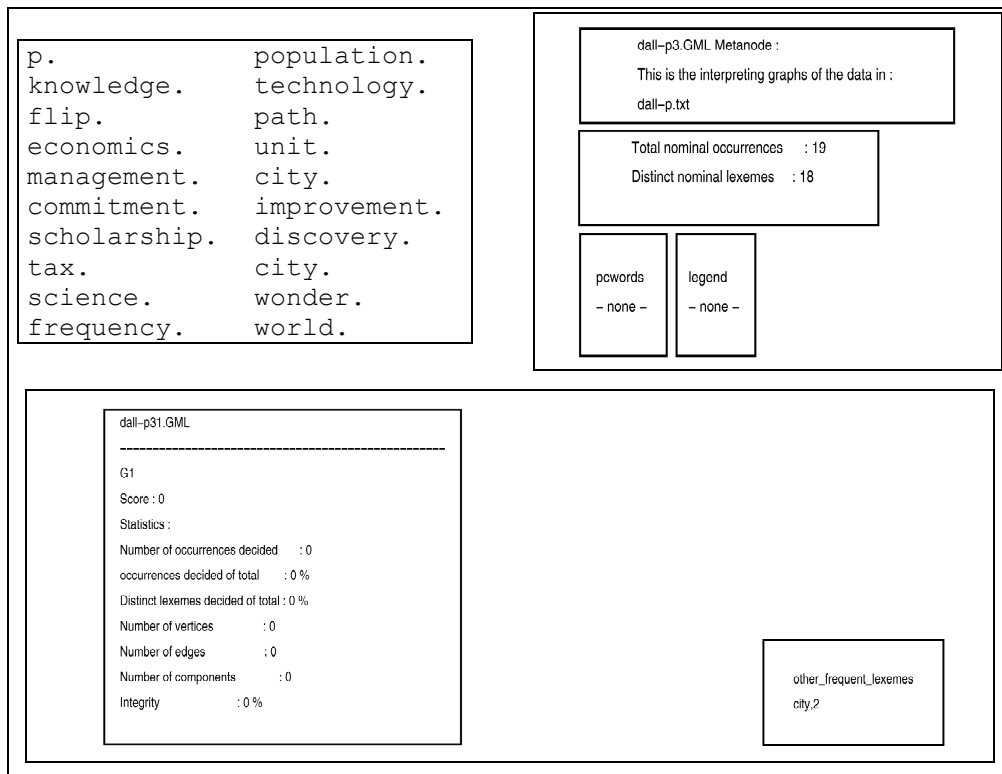


Figure A-1.4

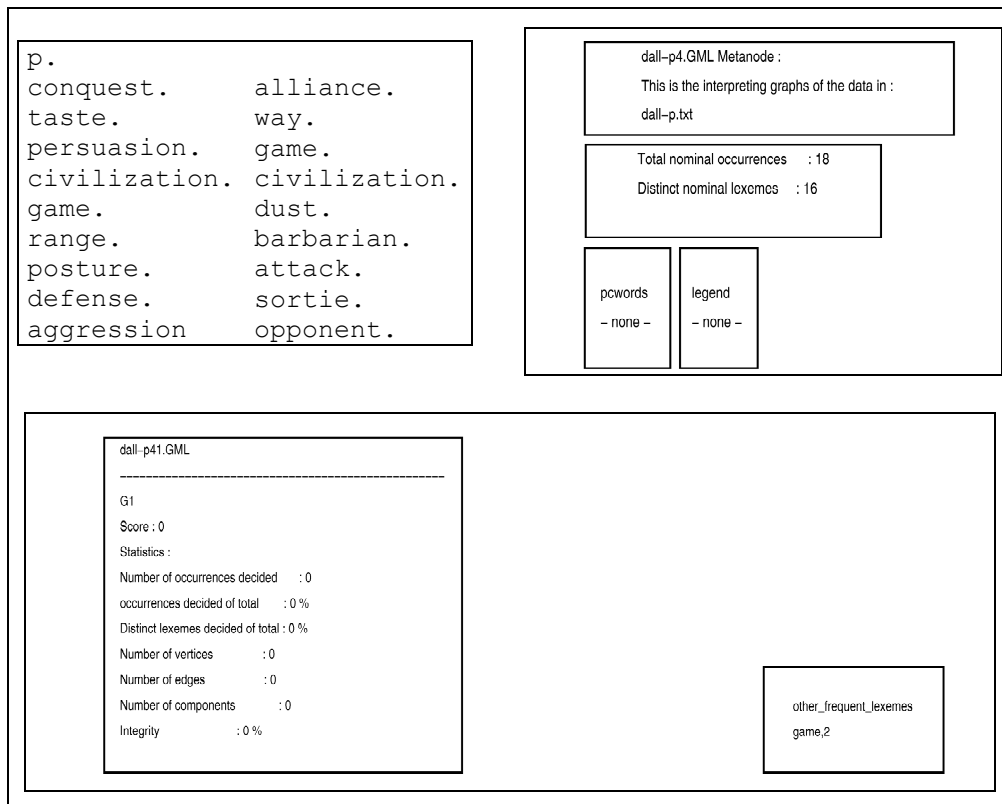


Figure A-1.5

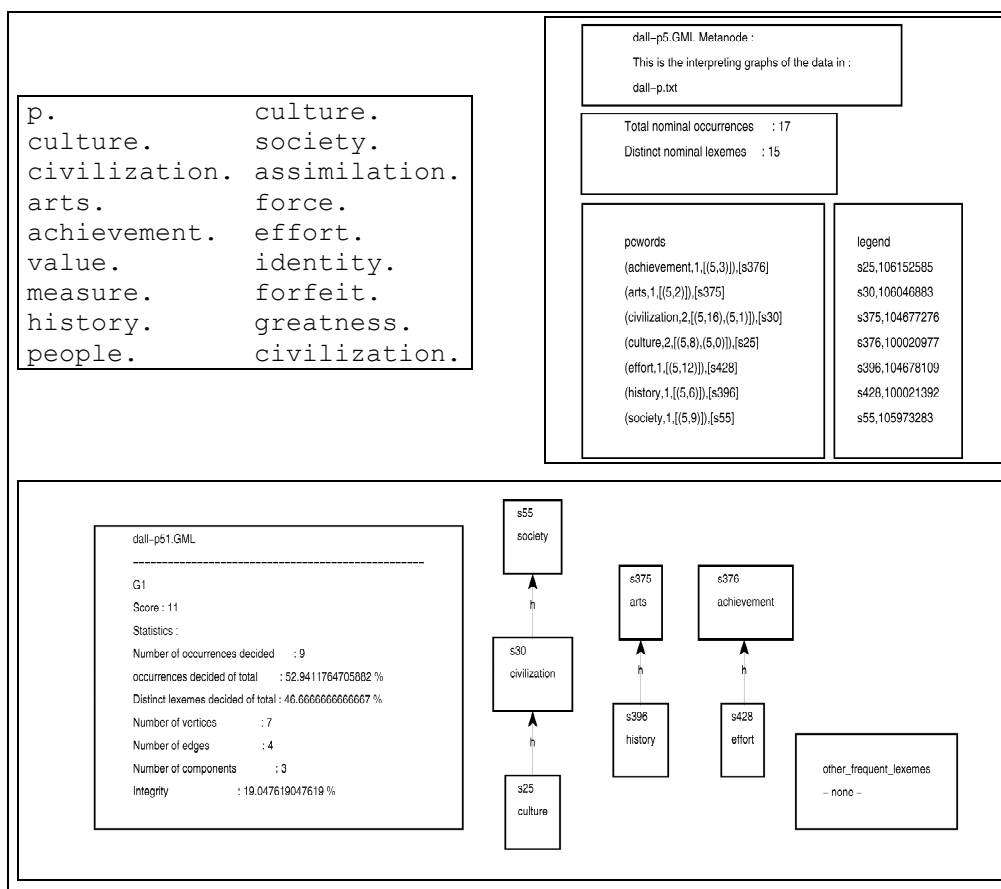


Figure A-1.6

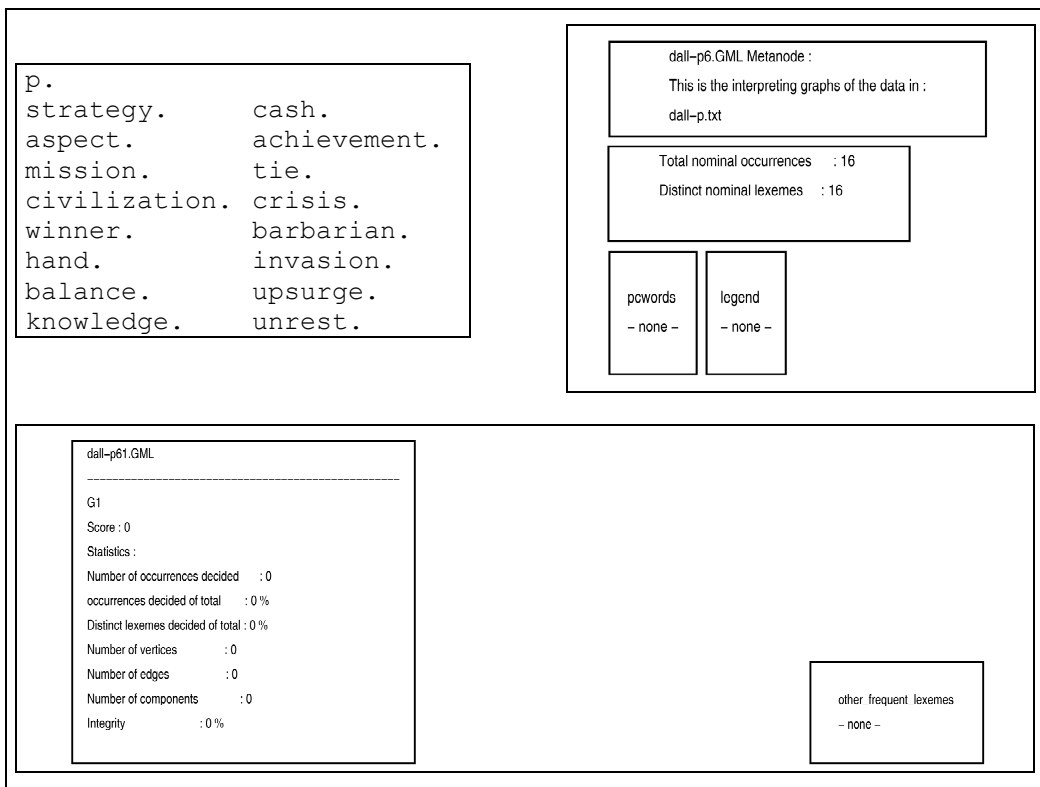


Figure A-1.7

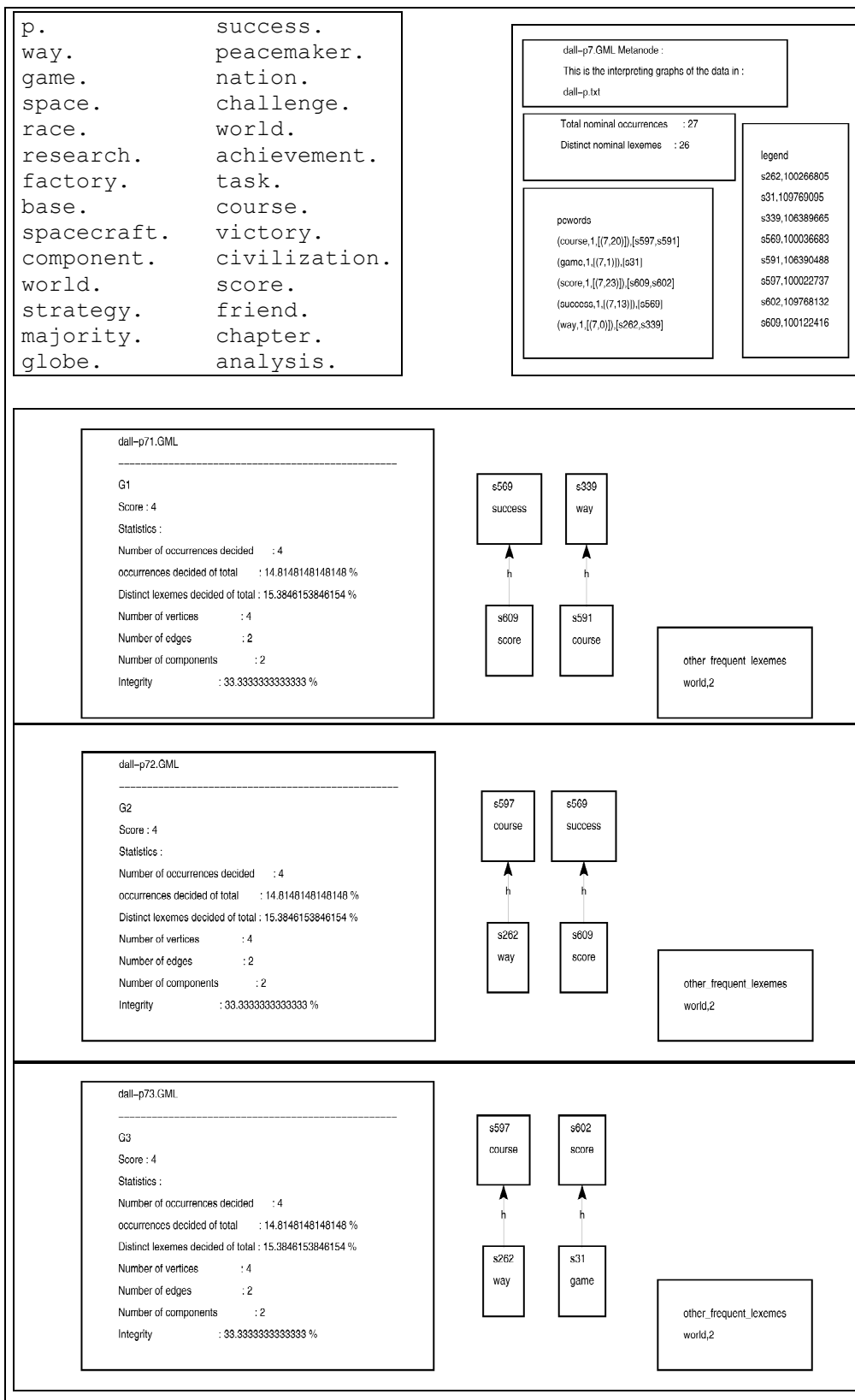


Figure A-1.8

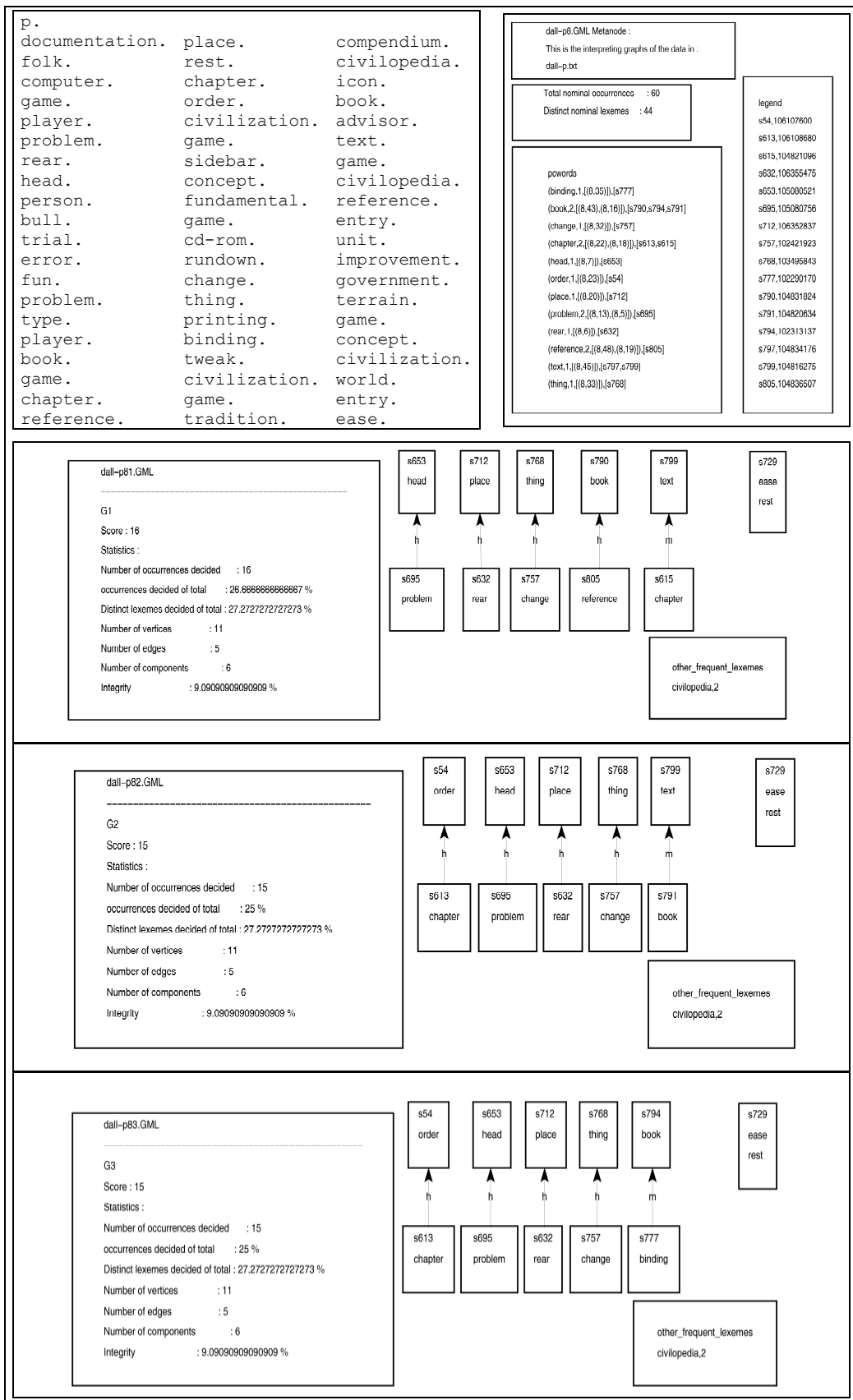
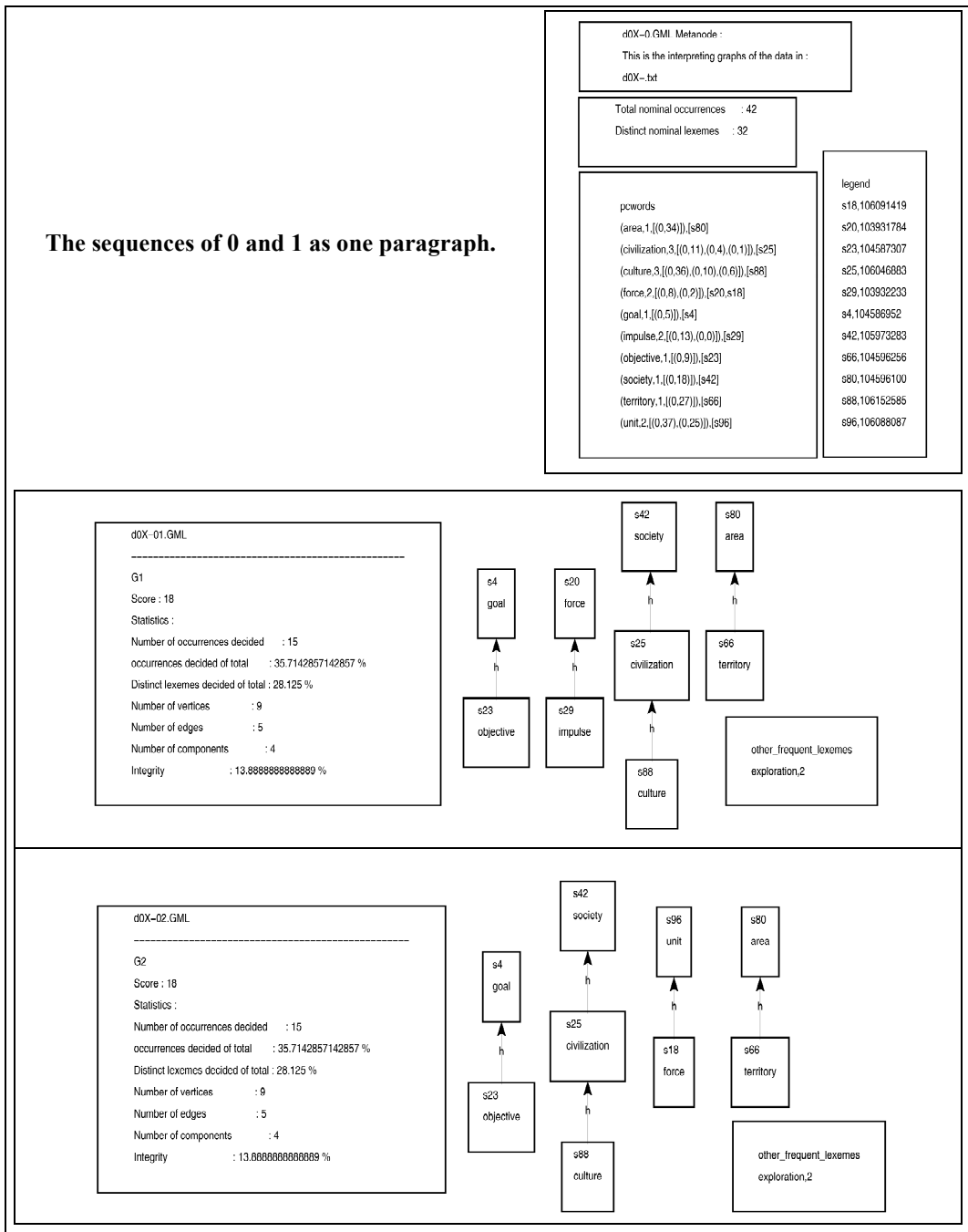


Figure A-1.9

A-2 Experiment 2 graphs



The sequences of 0 and 2 as one paragraph.

d0X-1.GML Metanode :
This is the interpreting graphs of the data in :
d0X-1.txt

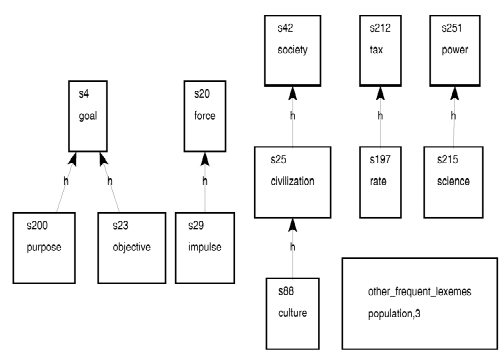
Total nominal occurrences : 48
Distinct nominal lexemes : 38

pcowords
(civilization,4,[(1,19),(1,11),(1,4),(1,1)],s25)
(culture,2,[(1,10),(1,6)],s86)
(force,2,[(1,8),(1,2)],s20)
(goal,1,[(1,5)],s4)
(impulse,2,[(1,13),(1,0)],s29)
(objective,1,[(1,9)],s23)
(power,1,[(1,43)],s251)
(purpose,1,[(1,20)],s200)
(rate,1,[(1,24)],s198,s197,s195)
(science,1,[(1,32)],s215)
(society,1,[(1,18)],s42)
(speed,1,[(1,27)],s205,s203)
(tax,2,[(1,31),(1,23)],s212)

legend
s195,110980504
s197,109585403
s198,103948875
s20,103931784
s200,104588033
s203,110979183
s205,103948579
s212,109580808
s215,104399665
s23,104587307
s25,106048883
s251,104340777
s29,103902203
s4,104586952
s42,106973283
s86,106152585

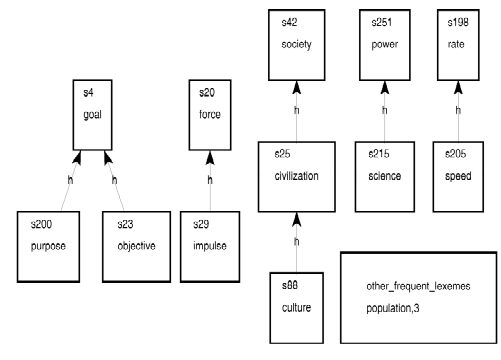
d0X-11.GML

G1
Score : 24
Statistics :
Number of occurrences decided : 19
occurrences decided of total : 39.58333333333333 %
Distinct lexemes decided of total : 31.5789473684211 %
Number of vertices : 12
Number of edges : 7
Number of components : 5
Integrity : 10.6060606060606 %



d0X-12.GML

G2
Score : 23
Statistics :
Number of occurrences decided : 18
occurrences decided of total : 37.5 %
Distinct lexemes decided of total : 31.5789473684211 %
Number of vertices : 12
Number of edges : 7
Number of components : 5
Integrity : 10.6060606060606 %



d0X-13.GML

G3
Score : 23
Statistics :
Number of occurrences decided : 18
occurrences decided of total : 37.5 %
Distinct lexemes decided of total : 31.5789473684211 %
Number of vertices : 12
Number of edges : 7
Number of components : 5
Integrity : 10.6060606060606 %

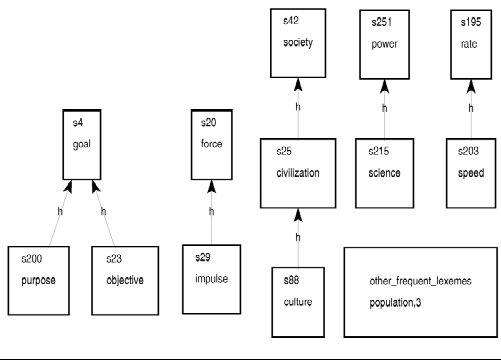
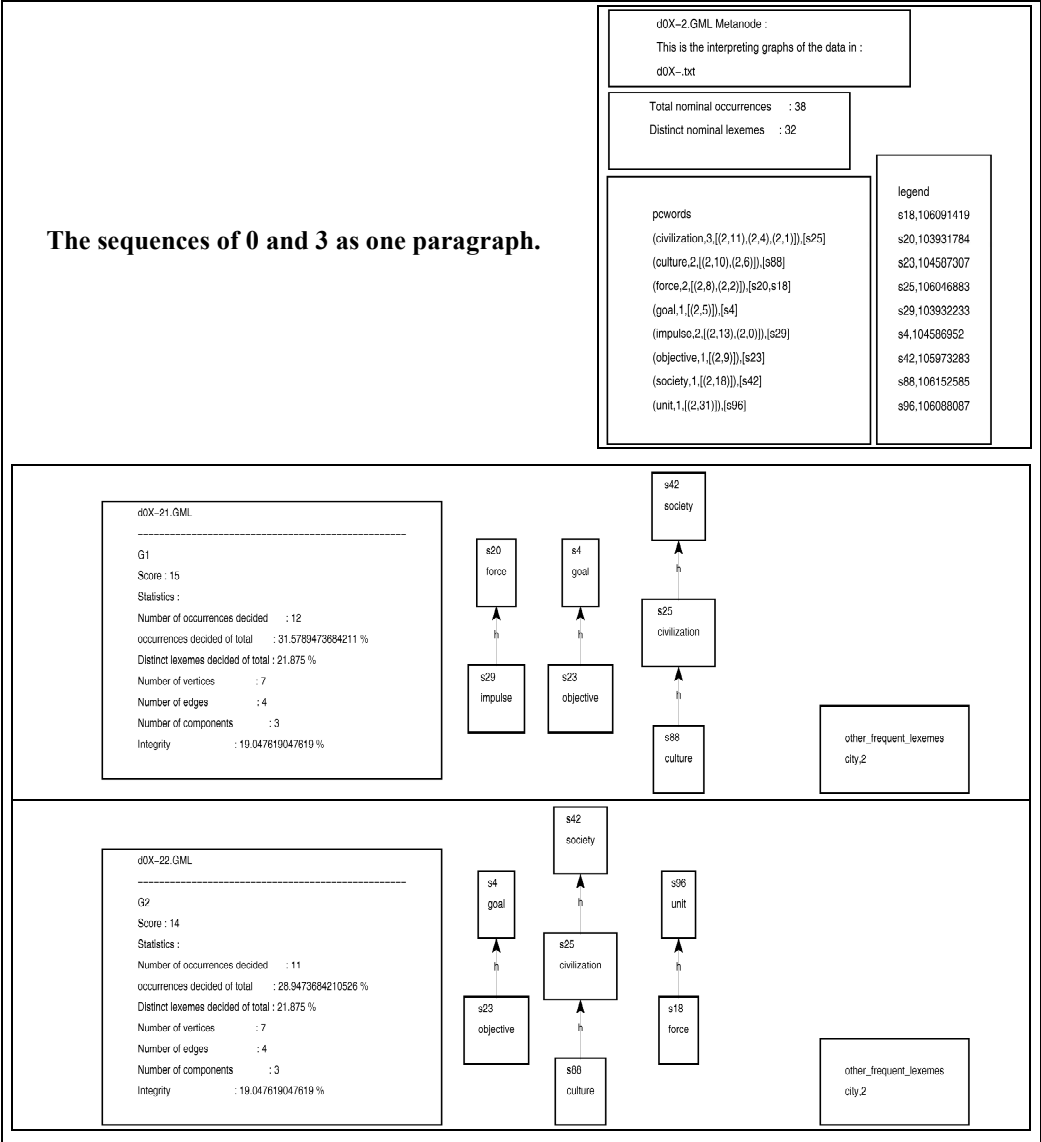


Figure A-2.2



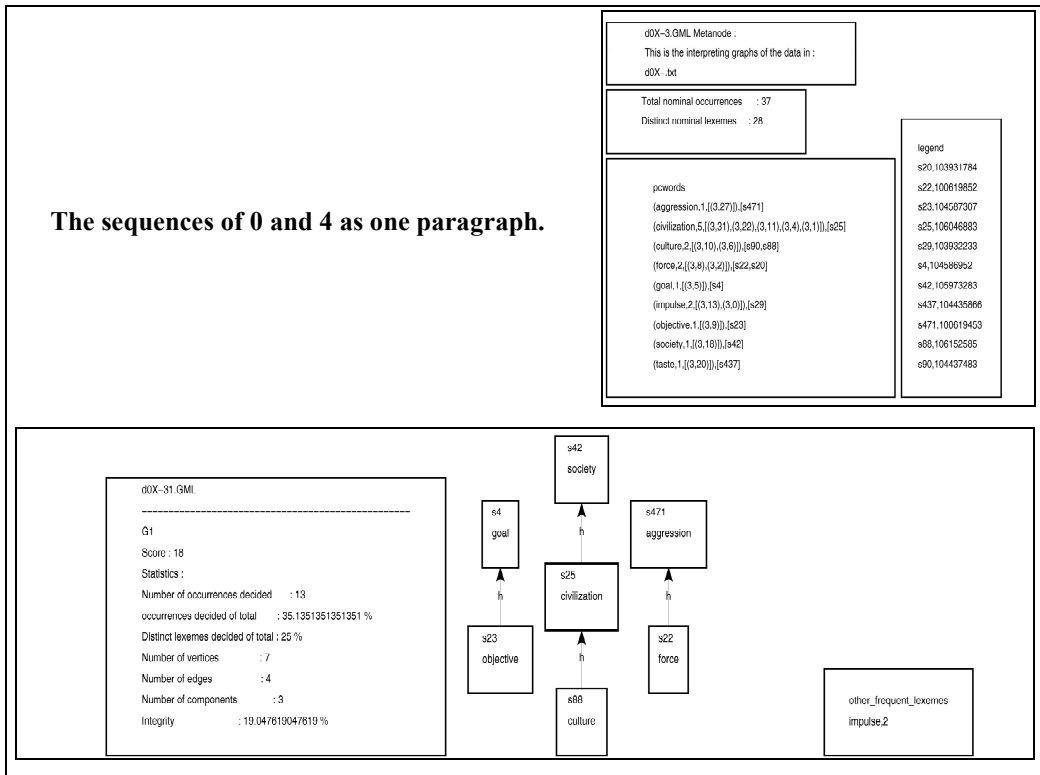


Figure A-2.4

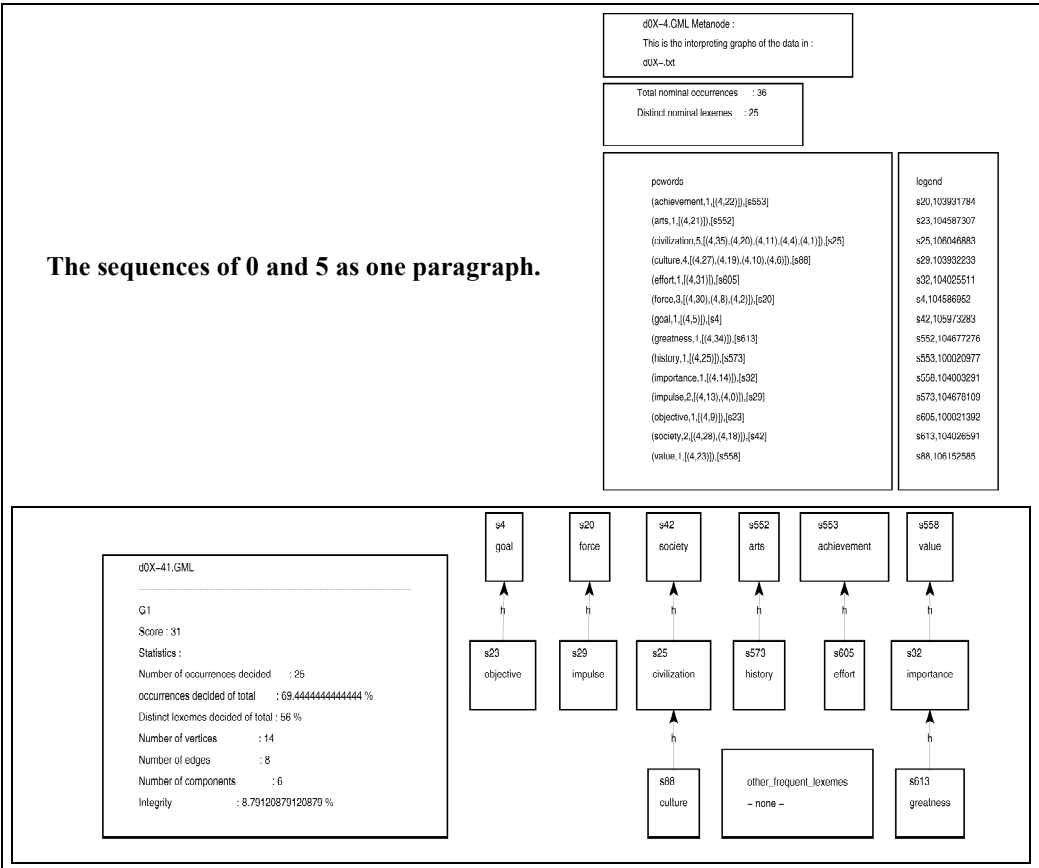


Figure A-2.5

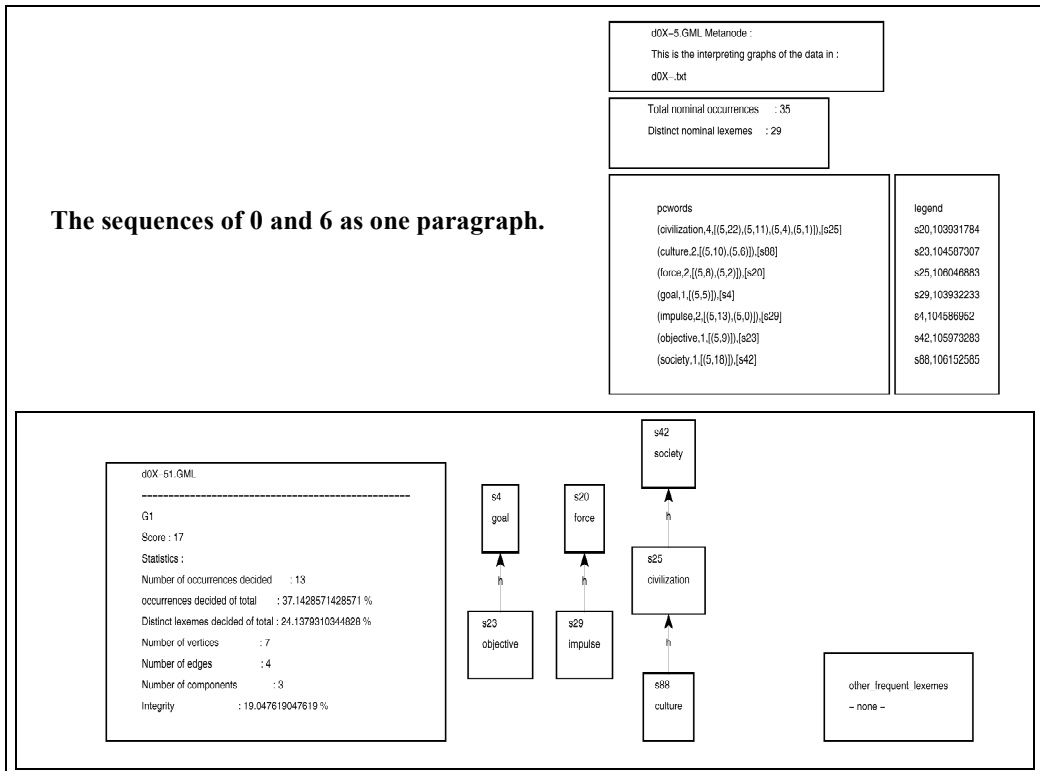
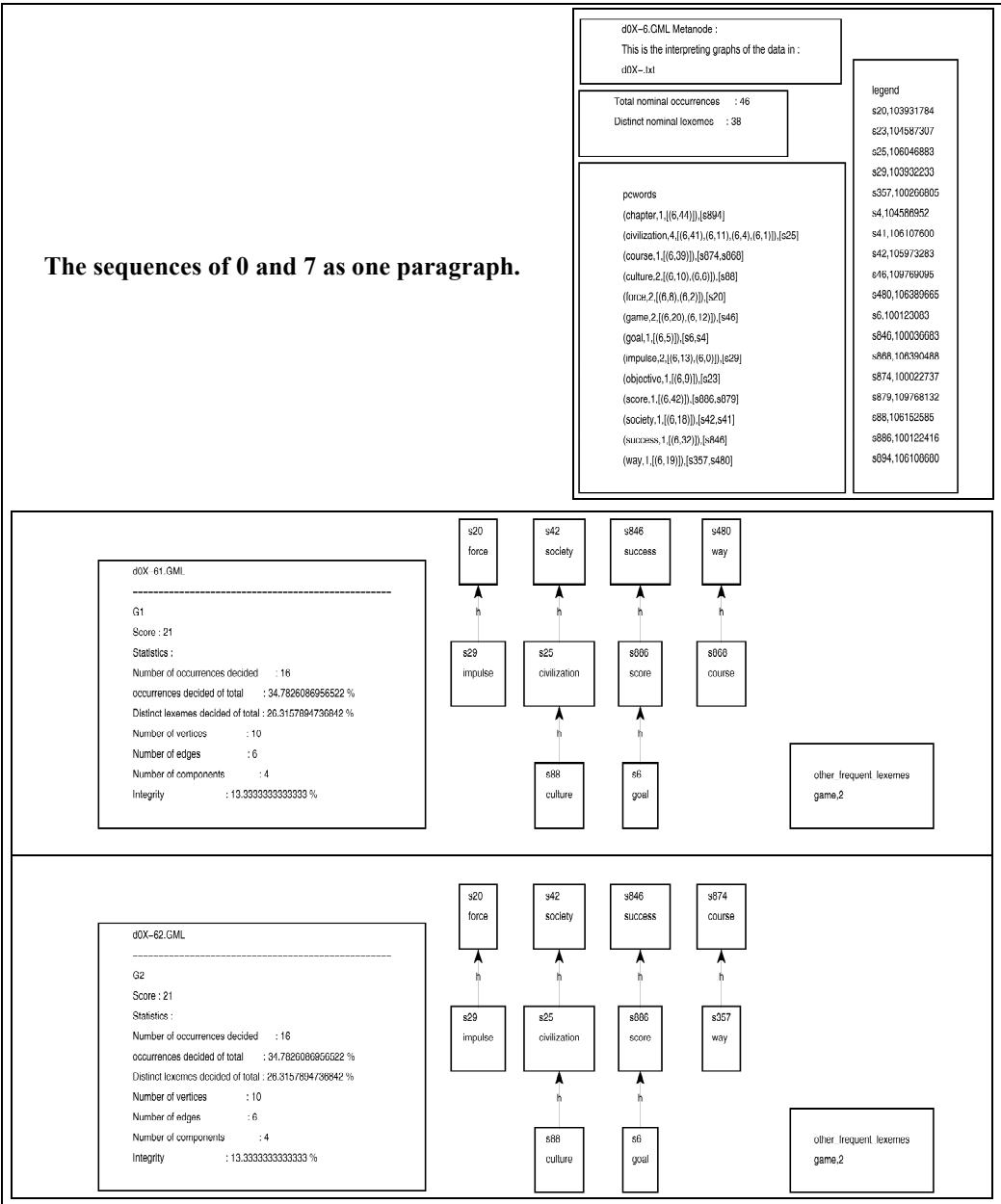
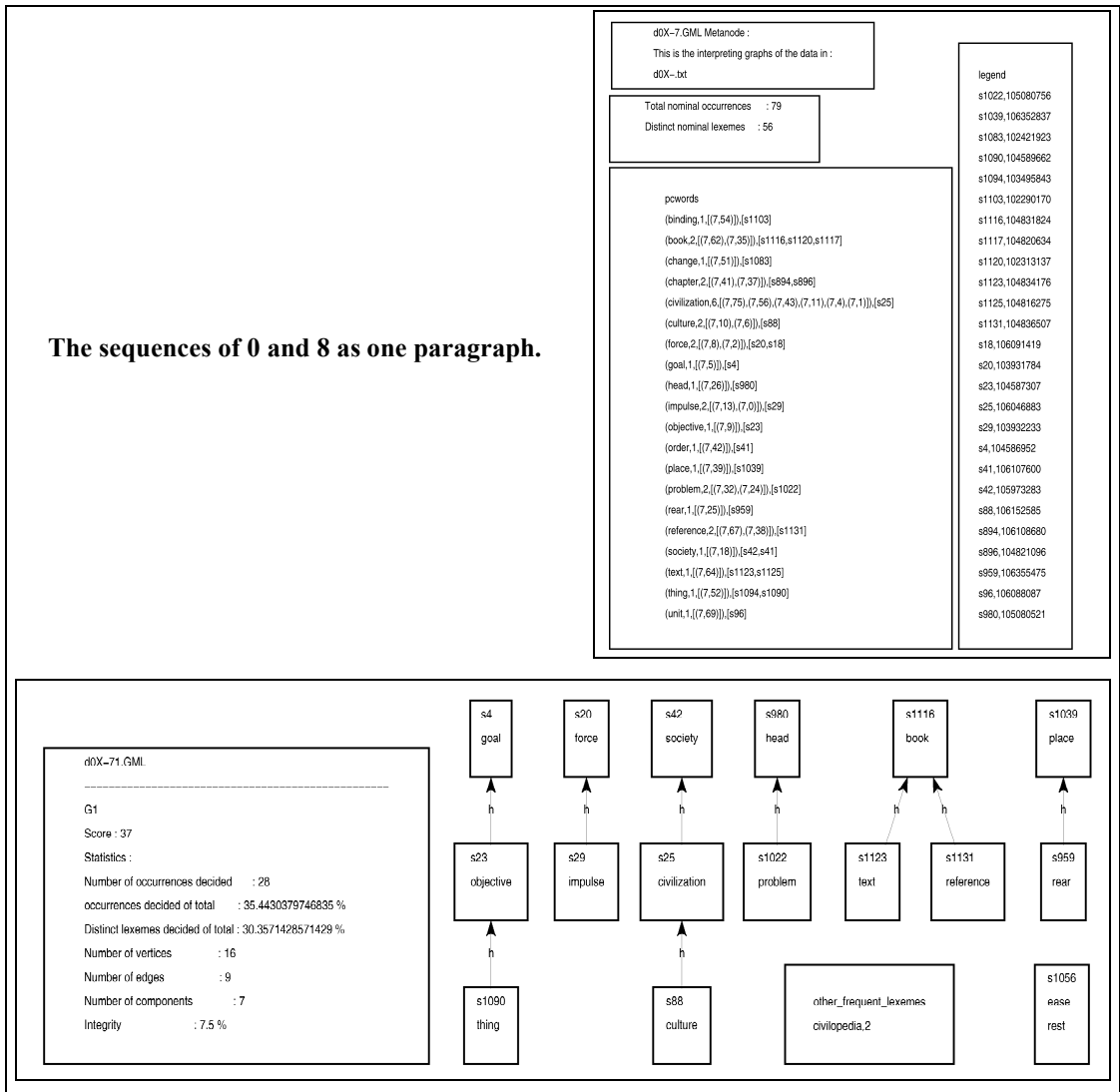


Figure A-2.6





A-3 Experiment 3 graph

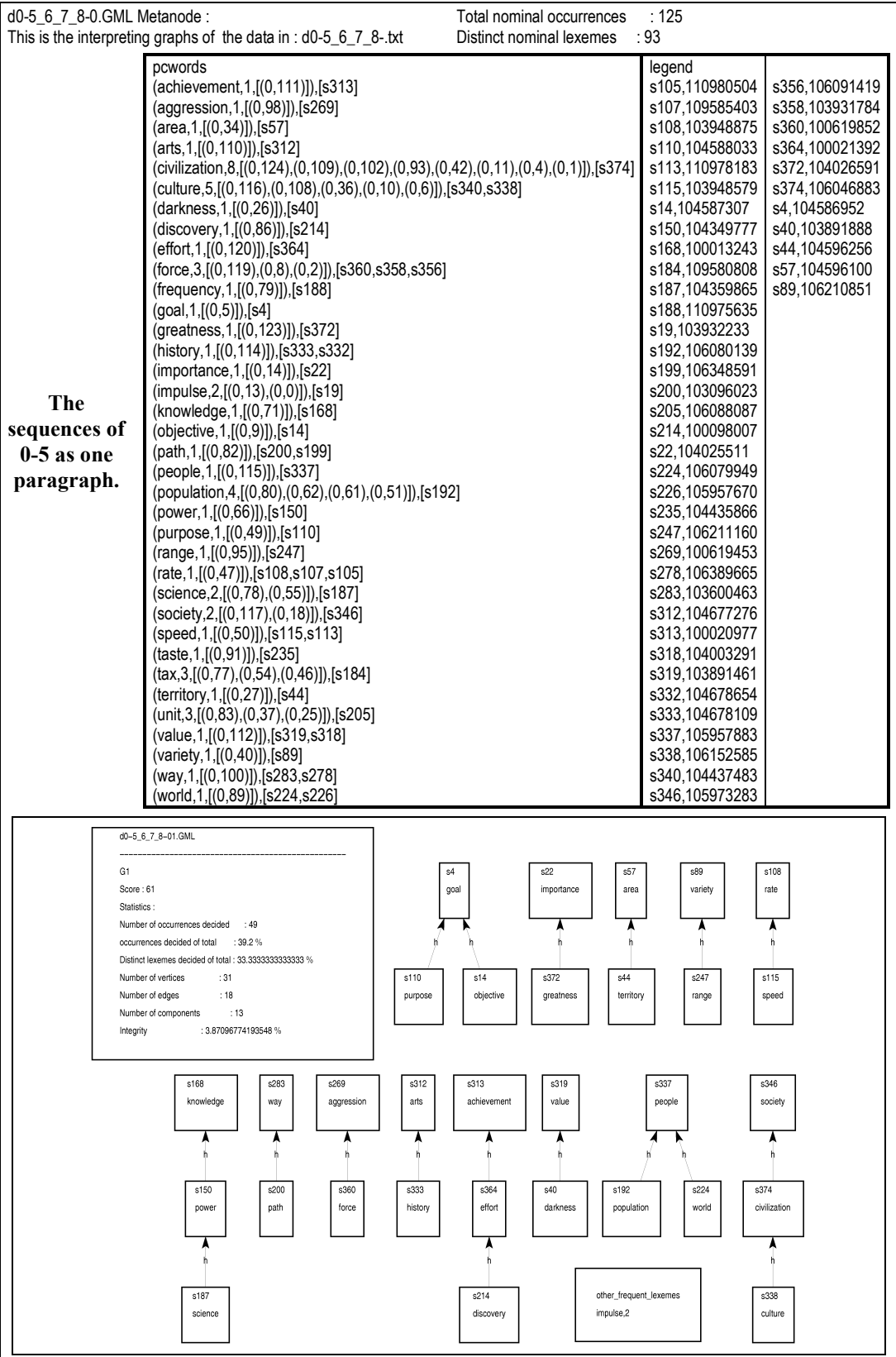


Figure A-3.1

A-4 Experiment 4 graph and glossary

The sequences of all paragraphs in one.		
dall-0.GML Metanode :	Total nominal occurrences : 228	
This is the interpreting graphs of the data in : dall-.txt	Distinct nominal lexemes : 160	
pcwords	legend	s444,100036683
(achievement,3,[(0,159),(0,134),(0,111)]),[s455]	s101,103175725	s447,106164129
(aggression,1,[(0,98)]),[s245]	s105,110980504	s448,106074189
(area,1,[(0,34)]),[s57,s59,s56]	s107,109585403	s449,106073126
(arts,1,[(0,110)]),[s267]	s108,103948875	s454,105375126
(aspect,1,[(0,126)]),[s330]	s109,104588033	s455,100020977
(binding,1,[(0,203)]),[s637]	s112,110978183	s458,106390488
(book,2,[(0,211),(0,184)]),[s653,s650,s654,s651]	s114,103948579	s464,100022737
(case,1,[(0,38)]),[s72,s70,s68]	s14,104587307	s468,109768132
(cash,2,[(0,133),(0,58)]),[s363]	s143,106078854	s475,100122416
(challenge,1,[(0,157)]),[s454]	s145,104349777	s476,100104367
(change,1,[(0,200)]),[s617,s614,s613]	s156,105374854	s478,107280379
(chapter,3,[(0,190),(0,186),(0,166)]),[s594,s596]	s162,106001253	s480,107060310
(civilization,13,[(0,224),(0,205),(0,192),(0,163),(0,128),(0,124),(0,109),(0,102),(0,93), (0,42),(0,11),(0,4),(0,1)]),[s693]	s178,109580808	s490,105975776
(component,1,[(0,149)]),[s431]	s181,104359865	s491,105961082
(conquest,1,[(0,90)]),[s207,s205]	s182,110975635	s496,106355475
(course,1,[(0,161)]),[s464,s458]	s186,106080139	s517,105080521
(culture,5,[(0,116),(0,108),(0,36),(0,10),(0,6)]),[s294,s292]	s19,103932233	s52,104504455
(darkness,1,[(0,26)]),[s40]	s193,106348591	s520,104345975
(defense,1,[(0,97)]),[s238]	s194,103096023	s532,100004123
(demand,1,[(0,69)]),[s156]	s195,100266805	s545,100775201
(discovery,1,[(0,86)]),[s198]	s198,100098007	s546,100508046
(effort,1,[(0,120)]),[s318,s316]	s205,100104245	s547,100507777
(empire,1,[(0,70)]),[s162]	s207,100037624	s559,105080756
(feature,1,[(0,33)]),[s52]	s211,104435866	s56,106268839
(folk,1,[(0,169)]),[s491,s490]	s22,104025511	s564,104497251
(force,3,[(0,119),(0,8),(0,2)]),[s314,s312,s310]	s223,106211160	s565,103556673
(frequency,1,[(0,79)]),[s182]	s238,100777065	s57,104596100
(friend,1,[(0,165)]),[s480,s478]	s245,100619453	s573,106368526
(game,13,[(0,222),(0,214),(0,206),(0,197),(0,193),(0,185),(0,171),(0,142),(0,101), (0,94),(0,22),(0,20),(0,12)]),[s684]	s266,107064973	s575,106352837
(goal,1,[(0,5)]),[s6,s4]	s267,104677276	s582,104346383
(government,1,[(0,220)]),[s679]	s272,104003291	s588,109947291
(greatness,1,[(0,123)]),[s326]	s273,103891461	s59,103995297
(hand,1,[(0,130)]),[s350]	s274,105418696	s594,106108680
(head,1,[(0,175)]),[s520,s517]	s286,104678654	s596,104821096
(history,1,[(0,114)]),[s287,s286]	s287,104678109	s6,100123083
(importance,1,[(0,14)]),[s22]	s290,105976176	s613,109642046
(impulse,2,[(0,13),(0,0)]),[s19]	s291,105957883	s614,109641836
(knowledge,2,[(0,132),(0,71)]),[s362]	s292,106152585	s617,102421923
(measure,1,[(0,113)]),[s274]	s294,104437483	s62,106073762
(nation,1,[(0,156)]),[s449,s447,s448]	s299,106107600	s624,104589662
(objective,1,[(0,9)]),[s14]	s300,105973283	s628,103495843
(opponent,1,[(0,107)]),[s266]	s310,106091419	s637,102290170
(order,1,[(0,191)]),[s299]	s312,103931784	s650,104831824
(path,1,[(0,82)]),[s195,s194,s193]	s314,100619852	s651,104820634
(people,1,[(0,115)]),[s290,s291]	s316,100503611	s653,102313477
(person,1,[(0,176)]),[s532]	s318,100021392	s654,102313137
(place,1,[(0,188)]),[s582,s401,s575,s573]	s326,104026591	s657,104834176
(population,4,[(0,80),(0,62),(0,61),(0,51)]),[s186]	s330,104504775	s659,104816275
(power,1,[(0,66)]),[s145,s143]	s340,107059664	s661,105417191
(problem,2,[(0,181),(0,173)]),[s559]	s350,104360663	s665,104836507
(production,1,[(0,44)]),[s101]	s362,100013243	s669,109946775
(purpose,1,[(0,49)]),[s109]	s363,109641408	s672,106088087
(range,1,[(0,95)]),[s223]	s386,106389665	s679,106000383
(rate,1,[(0,47)]),[s108,s107,s105]	s391,103600463	s68,107618819
(rear,1,[(0,174)]),[s496]	s396,106373003	s684,109769095
(reference,2,[(0,216),(0,187)]),[s665,s661]	s399,105134551	s693,106046883
(rest,1,[(0,189)]),[s588]	s4,104586952	s696,106079949
(science,2,[(0,78),(0,55)]),[s181]	s40,103891888	s698,105957670
(score,1,[(0,164)]),[s476,s475,s468]	s401,104817735	s70,107142349
(society,2,[(0,117),(0,18)]),[s300,s299]	s402,103193706	s72,105081510
(space,1,[(0,143)]),[s402,s401,s399,s396]	s431,109945970	s89,106210851
(speed,1,[(0,50)]),[s114,s112]	s44,104596256	s93,104496504

(success,1,[(0,154)],[s444,s340])
 (taste,1,[(0,91)],[s211])
 (tax,3,[(0,77),(0,54),(0,46)],[s178])
 (territory,1,[(0,27)],[s44])
 (text,1,[(0,213)],[s657,s659])
 (thing,1,[(0,201)],[s628,s624])
 (trial,1,[(0,178)],[s547,s546,s545])
 (tribe,1,[(0,35)],[s62])
 (type,1,[(0,182)],[s565,s564])
 (unit,4,[(0,218),(0,83),(0,37),(0,25)],[s672,s669])
 (value,1,[(0,112)],[s273,s272])
 (variety,1,[(0,40)],[s93,s89])
 (way,2,[(0,141),(0,100)],[s195,s391,s386])
 (winner,1,[(0,129)],[s340])
 (world,4,[(0,225),(0,158),(0,150),(0,89)],[s696,s698])

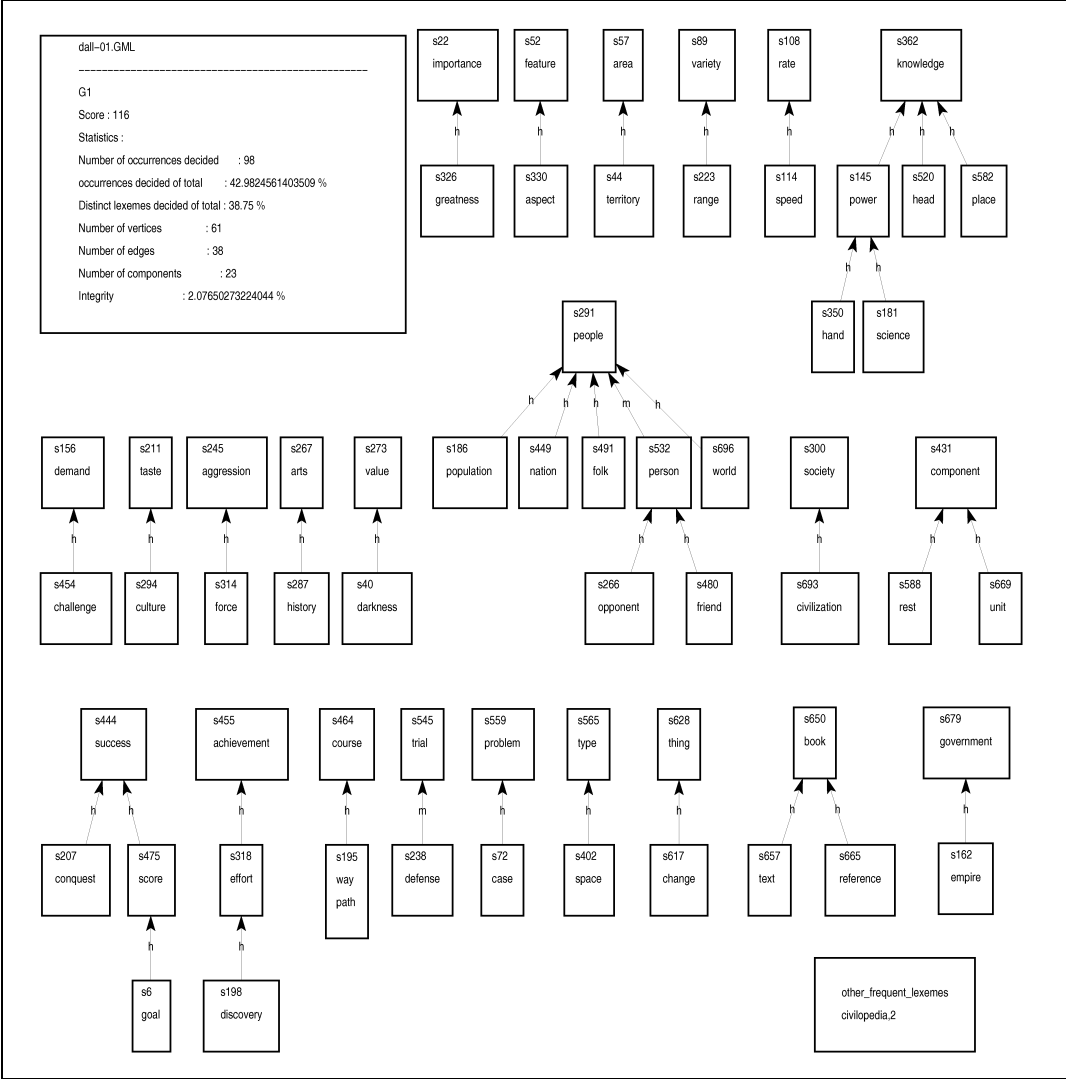


Figure A-4.1

The glossary of experiment 4

Glossary for datafile : dall-.txt

glossary

- s1,(an instinctive motive; "profound religious impulses")
- s2,(a strong restless desire; "why this urge to travel?")
- s3,(place where something (eg, a journey or race) ends)
- s4,(the state of affairs that a plan is intended to achieve and that (when achieved) terminates behavior intended to achieve it; "the ends justify the means")
- s5,(a place toward which players of a game try to advance a ball or puck in order to score points)
- s6,(a successful attempt at scoring; "the winning goal came with less than a minute left to play")
- s7,(an intricate network suggesting something that was formed by weaving or interweaving; "the trees cast a delicate web of shadows over the lawn")
- s8,(an intricately connected system of things or people; "a network of spies" or "a web of intrigue")
- s9,(a collection of internet sites that offer text and graphics and sound and animation resources through the hypertext transfer protocol)
- s10,(a fabric (especially a fabric in the process of being woven))
- s11,(an intricate trap that entangles or ensnares its victim)
- s12,(membrane connecting the toes of some aquatic birds and mammals)
- s13,(the flattened weblike part of a feather consisting of a series of barbs on either side of the shaft)
- s14,(the goal intended to be attained (and which is believed to be attainable); "the sole object of her trip was to see her children")
- s15,(the lens or system of lenses nearest the object being viewed)
- s16,(a sudden desire; "he bought it on impulse")
- s17,(electronics) a sharp transient wave in the normal electrical state (or a series of such transients); "the pulsations seemed to be coming from a star")
- s18,(the electrical discharge that travels along a nerve fiber; "they demonstrated the transmission of impulses from the cortex to the hypothalamus")
- s20,(the act of applying force suddenly; "the impulse knocked him over")
- s21,(a prominent status; "a person of importance")
- s22,(the quality of being important and worthy of note; "the importance of a well-balanced diet")
- s23,(a healthy state of well-being)
- s24,(the property of being flexible)
- s25,(the trait of being easily persuaded)
- s26,(the quality of being adaptable or variable; "he enjoyed the flexibility of his working arrangement")
- s27,(any new participant in some activity)
- s28,(young bird that has just fledged or become capable of flying)
- s29,(a systematic consideration; "he called for a careful exploration of the consequences")
- s30,(a careful systematic search)
- s31,(to travel for the purpose of discovery)
- s32,(the environmental condition)
- s33,(the area in which something exists or lives: "the country--the flat agricultural surround")
- s34,(a function such that for every element of one set there is a unique element of another set)
- s35,(a diagrammatic representation of the earth's surface (or part of it))
- s36,(absence of moral or spiritual values; "the powers of darkness")
- s37,(absence of light or illumination)
- s38,(an unilluminated area; "he moved off into the darkness")
- s39,(an unenlightened state; "he was in the dark concerning their intentions"; "his lectures dispelled the darkness")
- s40,(having a dark or somber color)
- s41,(a swarthy complexion)
- s42,(a territorial possession controlled by a ruling state)
- s43,(a region marked off for administrative or other purposes)
- s44,(an area of knowledge or interest; "his questions covered a lot of territory")
- s45,(a land mass that projects well above its surroundings; higher than a hill)
- s46,(a large natural stream of water (larger than a creek); "the river was navigable for 50 miles")
- s47,(land where grass or grasslike vegetation grows)
- s48,(land that is covered with trees and shrubs)
- s49,(the trees and other plants in a large densely wooded area)
- s50,(the principal (full-length) film in a program at a movie theater; "the feature tonight is 'Casablanca'")
- s51,(a special or prominent article in a newspaper or magazine; "they ran a feature on retirement planning")
- s52,(a prominent aspect of something: "the map showed roads and other features"; "generosity is one of his best characteristics")
- s53,(the characteristic parts of a person's face: eyes and nose and mouth and chin; "an expression of pleasure crossed his features"; "his lineaments were very regular")

The glossary of experiment 4

- s54**,(an article of merchandise that is displayed or advertised more than other articles)
- s55**,(a particular environment or walk of life; "his social sphere is limited"; "it was a closed area of employment"; "he's out of my orbit")
- s56**,(a particular geographical region of indefinite boundary (usually serving some special purpose or distinguished by its people or culture or geography); "it was a mountainous area"; "Bible country")
- s57**,(a subject of study; "it was his area of specialization"; "areas of interest include...")
- s58**,(a part of an animal that has a special function or is supplied by a given artery or nerve; "in the abdominal region")
- s59**,(the extent of a 2-dimensional surface enclosed within a boundary; "the area of a rectangle"; "it was about 500 square feet in area")
- s60**,(a part of a structure having some specific characteristic or function; "the spacious cooking area provided plenty of room for servants")
- s61**,(a social division of (usually preliterate) people)
- s62**,(a federation (as of American Indians))
- s63**,((biology) a taxonomic category between a genus and a subfamily)
- s64**,(group of people related by blood or marriage)
- s65**,(a specific state of mind that is temporary; "a case of the jitters")
- s66**,(a special set of circumstances; "in that event, the first possibility is excluded"; "it may rain in which case the picnic will be canceled")
- s67**,(the quantity contained in a case)
- s68**,(a person who is subjected to experimental or other observational procedures; someone who is an object of investigation; "the subjects for this investigation were selected randomly"; "the cases that we studied were drawn from two different communities")
- s69**,(a person of a specified kind (usually with many eccentricities); "a strange character"; "a friendly eccentric"; "the capable type"; "a mental case")
- s70**,(a person requiring professional services; "a typical case was the suburban housewife described by a marriage counselor")
- s71**,(an occurrence of something; "it was a case of bad judgment"; "another instance occurred yesterday"; "but there is always the famous example of the Smiths")
- s72**,(a problem requiring investigation; "Perry Mason solved the case of the missing heir")
- s73**,(a statement of facts and reasons used to support an argument; "he stated his case clearly")
- s74**,((nouns or pronouns or adjectives (often marked by inflection) related in some way to other words in a sentence)
- s75**,(the actual state of things; "that was not the case")
- s76**,(an enveloping structure or covering enclosing an animal or plant organ or part)
- s77**,(the outer covering or housing of something; "the clock has a walnut case")
- s78**,(the enclosing frame around a door or window opening; "the casings had rotted away and had to be replaced")
- s79**,(a cover for a pillow; "the burglar carried his loot in a pillowcase")
- s80**,(a glass container used to store and display items in a shop or museum or home)
- s81**,(a portable container for carrying several objects; "the musicians left their instrument cases backstage")
- s82**,((law) a comprehensive term for any proceeding in a court of law whereby an individual seeks a legal remedy; "the family brought suit against the landlord")
- s83**,(a place where two things come together; "Pittsburgh is located at the confluence of the Allegheny and Monongahela rivers")
- s84**,(a small informal social gathering; "there was an informal meeting in my livingroom")
- s85**,(a formally arranged gathering; "next year the meeting will be in Chicago")
- s86**,(a casual or unexpected convergence; "he still remembers their meeting in Paris"; "there was a brief encounter in the hallway")
- s87**,(the social act of assembling for some common purpose; "his meeting with the salesman was the high point of his day")
- s88**,(the act of joining together as one; "the merging of the two groups occurred quickly"; "there was no meeting of minds")
- s89**,(a collection containing a variety of sorts of things; "a great assortment of cars was on display"; "he had a variety of disorders")
- s90**,(a special kind of domesticated animals within a species; "he experimented on a particular breed of white rats"; "he created a new variety of sheep")
- s91**,(a show consisting of a series of short unrelated performances)
- s92**,(a specific kind of something; "a species of molecule"; "a species of villainy")
- s93**,(a category of things distinguished by some common characteristic or quality; "sculpture is a form of art"; "what kinds of desserts are there?")
- s94**,(a difference that is usually pleasant; "he goes to France for variety"; "it is a refreshing change to meet a woman mechanic")
- s95**,(noticeable heterogeneity; "a diversity of possibilities"; "the range and variety of his work is amazing")
- s96**,(a hostile disagreement face-to-face)
- s97**,(a minor short-term fight)
- s98**,(a casual meeting with a person of thing)
- s99**,(the quality of being intricate and compounded; "he enjoyed the complexity of modern computers")
- s100**,(a presentation for the stage or screen or radio or television; "have you seen the new production of Hamlet?")
- s101**,(the amount of an artifact that has been produced by someone or some process; "they improve their product every year"; "they export most of their agricultural production")
- s102**,((economics) manufacturing or mining or growing something (usually in large quantities) for sale; "he introduced more efficient methods of production")
- s103**,(the act or process of producing something; "Shakespeare's production of poetry was enormous"; "the production of white blood cells")

The glossary of experiment 4

- s104.**((law) the act of exhibiting in a court of law; "the appellate court demanded the production of all documents")
- s105.**(a magnitude or frequency relative to a time unit; "they traveled at a rate of 55 miles per hour"; "the rate of change was faster than expected")
- s106.**(amount of a charge or payment relative to some basis; "a 10-minute phone call at that rate would cost \$5")
- s107.**((British) a local tax on property (usually used in the plural))
- s108.**(the relative speed of progress or change; "he lived at a fast pace"; "he works at a great rate"; "the pace of events accelerated")
- s109.**(an anticipated outcome that is intended or guides your planned actions; "his intent was to provide a new translation"; "it was created with the conscious aim of answering immediate needs"; "he made no secret of his designs")
- s110.**(what something is used for; "the function of an auger is to bore holes"; "ballet is beautiful but what use is it?")
- s111.**(the quality of being determined to do or achieve something; "his determination showed in his every movement"; "he is a man of purpose")
- s112.**(distance travelled per unit time)
- s113.**(the ratio of the focal length to the diameter of a (camera) lens system)
- s114.**(a rate (usually rapid) at which something happens; "the project advanced with gratifying speed")
- s115.**(a central nervous system stimulant that increases energy and decreases appetite; used to treat narcolepsy and some forms of depression)
- s116.**(changing location rapidly)
- s117.**(articles of commerce)
- s118.**(a combat between two mounted knights tilting against each other with blunted lances)
- s119.**(a dispute where there is strong disagreement; "they were involved in a violent argument") **s120.**(a slight but noticeable partiality; "the court's tilt toward conservative rulings")
- s121.**(the property possessed by a line or surface that departs from the vertical; "the tower had a pronounced tilt"; "the ship developed a list to starboard"; "he walked with a heavy inclination to the right")
- s122.**(pitching dangerously to one side)
- s123.**(the system of production and distribution and consumption)
- s124.**(the efficient use of resources; "economy of effort")
- s125.**(frugality in the expenditure of money or resources; "the Scots are famous for their economy")
- s126.**(an act of economizing; reduction in cost; "it was a small economy to walk to work every day" or "there was a saving of 50 cents")
- s127.**(a large unpleasant woman)
- s128.**(female of domestic cattle: "moo-cow" is a child's term")
- s129.**(mature female of mammals of which the male is called 'bull')
- s130.**(state of well-being characterized by emotions ranging from contentment to intense joy)
- s131.**(emotions experienced when in a state of well-being)
- s132.**(an activity that entertains)
- s133.**(the state of being present; current existence; "he tested for the presence of radon")
- s134.**(an invisible spiritual being felt to be nearby)
- s135.**(the immediate proximity of someone or something; "she blushed in his presence"; "he sensed the presence of danger"; "he was well behaved in front of company")
- s136.**(the impression that something is present; "he felt the presence of an evil force")
- s137.**(dignified manner or conduct)
- s138.**(the act of being present)
- s139.**((of a government or government official) holding an office means being in power; "being in office already gives a candidate a great advantage"; "during his first year in power")
- s140.**((physics) the rate of doing work; measured in watts (= joules/second))
- s141.**(one possessing or exercising power or influence or authority: "the mysterious presence of an evil power"; "may the force be with you"; "the forces of evil")
- s142.**(a very wealthy or powerful businessman: "an oil baron")
- s143.**(a state powerful enough to influence events throughout the world)
- s144.**(a mathematical notation indicating the number of times a quantity is multiplied by itself)
- s145.**(possession of the qualities (especially mental qualities) required to do something or get something done; "danger heightened his powers of discrimination")
- s146.**(possession of controlling influence; "the deterrent power of nuclear weapons"; "the power of his love saved her")
- s147.**(physical strength)
- s148.**(wealth as evidenced by sumptuous living)
- s149.**(the quality possessed by something that is excessively expensive)
- s150.**(something that is an indulgence rather than a necessity)
- s151.**(available source of wealth; a new or reserve supply that can be drawn upon when needed)
- s152.**(the ability to deal resourcefully with unusual problems; "a man of resource")
- s153.**(a source of aid or support that may be drawn upon when needed: "the local library is a valuable resource")
- s154.**(a condition requiring relief; "she satiated his need for affection"; "God has no need of men to accomplish His work"; "there is a demand for jobs")
- s155.**(the ability and desire to purchase goods and services; "the automobile reduced the demand for buggywhips"; "the demand exceeded the supply")
- s156.**(an urgent or peremptory request; "his demands for attention were unceasing")
- s157.**(required activity; "the requirements of his work affected his health"; "there were many demands on his time")
- s158.**(the act of demanding; "the kidnapper's exorbitant demands for money")
- s159.**(the domain ruled by an emperor or empress)
- s160.**(a monarchy with an emperor as head of state)

The glossary of experiment 4

- s161**,(a group of companies run as a single organization)
s162,(a group of countries under a single authority: "the Roman empire")
s163,(hot or cold alcoholic drink containing beaten egg)
s164,(the act of flipping a coin)
s165,(a dive in which the diver somersaults before entering the water)
s166,(an acrobatic feat in which the feet roll over the head (either forward or backward) and return)
s167,((sports) the act of throwing the ball to another member of your team; "the pass was fumbled")
s168,(the branch of social science that deals with the production and distribution and consumption of goods and services and their management)
s169,(those in charge of running a business)
s170,(the act of managing something; "he was given overall management of the program"; "is the direction of the economy a function of government?")
s171,(a message that makes a pledge)
s172,(the trait of sincere and steadfast fixity of purpose; "a man of energy and commitment")
s173,(an engagement by contract involving financial obligation; "his business commitments took him to London")
s174,(the act of binding yourself (intellectually or emotionally) to a course of action; "his long commitment to public service"; "they felt no loyalty to a losing team")
s175,(the official act of consigning a person to a prison (or mental hospital))
s176,(financial aid provided to a student on the basis of academic merit)
s177,(profound knowledge)
s178,(charge against a citizen's person or property or activity for the support of government)
s179,(a particular branch of scientific knowledge; "the science of genetics")
s180,(any domain of knowledge accumulated by systematic study and organized by general principles; "mathematics is important for science")
s19,(an impelling force or strength; "the car's momentum carried it off the road")
s181,(ability to produce solutions in some problem domain; "the skill of a well-trained boxer"; "the science of pugilism")
s182,(the number of occurrences within a given time period (usually 1 second); "the frequency of modulation was 40 cycles per second")
s183,(the ratio of the number of observations in a statistical category to the total number of observations)
s184,(the number of observations in a given statistical category)
s185,(the number of inhabitants in a given place (country or city etc.); "people come and go, but the population of this town has remained approximately constant for the past decade")
s186,(the people who inhabit a territory or state; "the population seemd to be well fed and clothed")
s187,(a group of organisms of the same species populating a given area; "they hired hunters to keep down the deer population")
s188,((statistics) the entire aggregation of items from which samples can be drawn; "it is an estimate of the mean of the population")
s189,(the act of populating (causing to live in a place); "he deplored the population of colonies with convicted criminals")
s190,(the discipline dealing with the art or science of applying scientific knowledge to practical problems; "he had trouble deciding which branch of engineering to study")
s191,(the practical application of science to commerce or industry)
s192,(a line or route along which something travels or moves: "the hurricane demolished houses in its path"; "the track of an animal"; "the course of the river")
s193,(an established line of travel or access)
s194,(a way especially designed for a particular use)
s195,(a course of conduct; "the path of virtue"; "we went our separate ways"; "our paths in life led us apart"; "genius usually follows a revolutionary path")
s196,(something that is discovered)
s197,(a productive insight)
s198,(the act of discovering something)
s199,(an incorporated administrative district established by state charter; "the city raised the tax rate")
s200,(a large and densely populated urban area; may include several independent administrative districts; "Ancient Troy was a great city")
s201,(people living in a large densely populated municipality; "the city voted for Republicans in 1994")
s202,(the feeling aroused by something strange and surprising)
s203,(something that causes feelings of wonder; "the wonders of modern science")
s204,(a state in which you want to learn more about something)
s205,(an act of winning the love of someone)
s206,(the act of conquering)
s207,(success in mastering something difficult; "the conquest of space")
s208,(a small amount eaten or drunk; "take a taste--you'll like it")
s209,(a strong liking; "my own preference is for good literature"; "the Irish have a penchant for blarney"; "martinis are an acquired taste")
s210,(a brief experience of something; "he got a taste of life on the wild side"; "she enjoyed her brief taste of independence")
s211,(delicate discrimination (especially of aesthetic values); "arrogance and lack of taste contributed to his rapid success"; "to ask at that particular time was the ultimate in bad taste")
s212,(the sensation that results when taste buds in the tongue and throat convey information about the chemical composition of a soluble stimulus; "the candy left him with a bad taste"; "the melon had a delicious taste")
s213,(the faculty of taste; "his cold deprived him of his sense of taste")
s214,(distinguishing a taste by means of the taste buds; "he loved the smell and taste of fresh bread"; "a wine tasting")

The glossary of experiment 4

- s215,(the act of persuading (or attempting to persuade))
s216,(a personal belief that is not founded on proof or certainty; "my opinion differs from yours"; "what are your thoughts on Haiti?")
s217,(inducement by argument or reasoning or entreaty)
s218,(changing a person's beliefs by argument or reasoning or entreaty)
s219,(a series of hills or mountains; "the valley was between two ranges of hills"; "the plains lay just beyond the mountain range")
s220,(the limits of the values a function can take; "the range of this function is the interval from 0 to 1")
s221,(a large tract of grassy open land on which livestock can graze; "they used to drive the cattle across the open range every spring"; "he dreamed of a home on the range")
s222,(the limits within which something can be effective; "he was beyond the range of their fire")
s223,(a variety of different things or activities; "he answered a range of questions"; "he was impressed by the range and diversity of the collection")
s224,(the limit of capability; "within the compass of education")
s225,(an area in which something acts or operates or has power or control: "the range of a supersonic jet"; "the ambit of municipal legislation"; "within the compass of this article"; within the scope of an investigation"; "outside the reach of the law"; "in the political orbit of a world power")
s226,(a kitchen appliance used for cooking food; "dinner was already on the stove")
s227,(a place for shooting (firing or driving) projectiles of various kinds; "the army maintains a missile range in the desert"; "any good golf club will have a range")
s228,(a rationalized mental attitude)
s229,(position or arrangement of the body and its limbs; "he assumed an attitude of surrender")
s230,(characteristic way of bearing one's body: "stood with good posture")
s231,(an unconscious process that tries to reduce the anxiety associated with instinctive desires)
s232,(the team that is trying to prevent the other team from scoring; "his teams are always good on defense")
s233,(the defendant and his legal advisors collectively; "the defense called for a mistrial")
s234,(an organization of defenders that provides resistance against attack; "he joined the defense against invasion")
s235,(the speech act of answering an attack on your assertions; "his refutation of the charges was short and persuasive"; "in defense he said the other man started it")
s236,(the justification for some act or belief; "he offered a persuasive defense of the theory")
s237,(a structure used for defense; "the artillery battered down the defenses")
s238,(a defendant's answer or plea denying the truth of the charges against him; "he gave evidence for the defense")
s239,(military action or resources protecting a country against potential enemies; "they died in the defense of Stalingrad"; "they were developed for the defense program")
s240,(protection from harm; "sanitation is the best defense against disease")
s241,(a feeling of hostility that arouses thoughts of attack)
s242,(a disposition to behave aggressively)
s243,(the act of initiating hostilities)
s244,(deliberately unfriendly behavior)
s245,(violent action that is hostile and usually unprovoked)
s246,(the state of being allied or confederated)
s247,(a connection based on kinship or marriage or common interest: "the shifting alliances within a large family"; "their friendship constitutes a powerful bond between them")
s248,(an organization of people (or countries) involved in a pact or treaty)
s249,(a formal agreement establishing an association or alliance between nations or other groups to achieve a particular aim)
s250,(the act of forming an alliance or confederation)
s251,(the remains of something that has been destroyed or broken up)
s252,(free microscopic particles of solid material; "astronomers say that the empty space between planets actually contains measurable amounts of dust")
s253,(fine powdery material such as dry earth or pollen that can be blown about in the air; "the furniture was covered with dust")
s254,(a sudden occurrence of an uncontrollable condition; "an attack of diarrhea")
s255,(intense adverse criticism; "Clinton directed his fire at Jesse Helms")
s256,(turning your attention to a problem or a job etc.; "his attack on the problem was misguided"; "he did not make a direct assault on her affections")
s257,(the beginning of an offensive; "the attack began at dawn")
s258,(a formulation adopted in tackling a problem; "his approach to every problem is to draw up a list of pros and cons")
s259,(an assault on someone; "they made an attempt on his life")
s260,(an offensive move in a sport or game; "they won the game with a 10-hit attack in the 9th inning")
s261,(a decisive manner of beginning a musical tone or phrase)
s262,(a military action in which besieged troops burst forth from their position)
s263,(an operational flight by a single aircraft (as in a military operation))
s264,(a contestant that you are matched against)
s265,(an armed adversary (especially a member of an opposing military force); "a soldier must be prepared to kill his enemies")
s266,(someone who offers opposition)
s267,(studies intended to provide general knowledge and intellectual skills (rather than occupational or professional skills); "the college of arts and sciences")
s268,(music) the relative duration of a musical note)
s269,(the amount (of money or goods or services) that is considered to be a fair equivalent for something else; "he tried to

The glossary of experiment 4

estimate the value of the produce at normal prices")

s270.(an ideal accepted by some individual or group; "he has old-fashioned values")

s271.(a numerical quantity measured or assigned or computed; "the value assigned was 16 milliseconds")

s272.(the quality (positive or negative) that renders something desirable or valuable; "the Shakespearean Shylock is of dubious value in the modern world")

s273.(relative darkness or lightness of a color: "I establish the colors and principal values by organizing the painting into three values--dark, medium...and light"-Joe Hing Lowe)

s274.(a basis for comparison; a reference point against which other things can be evaluated; "they set the measure for all subsequent work")

s275.((prosody) the accent in a metrical foot of verse)

s276.(notation for a repeating pattern of musical beats; written followed by a vertical bar)

s277.(a statute in draft before it becomes law; "they held a public hearing on the bill")

s278.(magnitude as determined by measurement or calculation)

s279.(a measuring instrument having a sequence of marks at regular intervals; used as a reference in making measurements)

s280.(the act or process of measuring; "he made a careful measurement"; "his mental measurings proved remarkably accurate")

s281.(any maneuver made as part of progress toward a goal; "the police took steps to reduce crime")

s282.(how much there is of something that you can measure)

s283.(the continuum of events occurring in succession leading from the past to the present and even into the future: "all of human history")

s284.(the aggregate of past events: "a critical time in the school's history")

s285.(a record or narrative description of past events: "a history of France"; "he gave an inaccurate account of the plot to kill the president"; "the story of exposure to lead")

s286.(all that is remembered of the past as preserved in writing; a body of knowledge: "the dawn of recorded history"; "from the beginning of history")

s287.(the discipline that records and interprets past events involving human beings: "he teaches Medieval history"; "history takes the long view")

s288.(the common people generally; "separate the warriors from the mass"; "power to the people")

s289.(the body of citizens of a state or country; "the Spanish people")

s290.(members of a family line; "his people have been farmers for generations"; "are your people still alive?")

s291.((plural) any group of human beings (men or women or children) collectively; "old people"; "there were at least 200 people in the audience")

s292.(a particular civilization at a particular stage)

s293.(all the knowledge and values shared by a society)

s294.(the tastes in art and manners that are favored by a social group)

s295.((biology) the growing of microorganisms in a nutrient medium (such as gelatin or agar); "the culture of cells in a Petri dish")

s296.(the raising of plants or animals: "the culture of oysters")

s297.(the state of being with someone; "he missed their company"; "he enjoyed the society of his friends")

s298.(the fashionable elite)

s299.(a formal association of people with similar interests; "he joined a golf club"; "they formed a small lunch society"; "men from the fraternal order will staff the soup kitchen today")

s300.(an extended social group having a distinctive cultural and economic organization)

s301.(the state of being assimilated)

s302.(a linguistic process by which a sound becomes similar to an adjacent sound)

s303.(the process of absorbing nutrients into the body after digestion)

s304.(the process of assimilating new ideas into an existing cognitive structure)

s305.(in the theories of Jean Piaget: the application of a general schema to a particular instance)

s306.(the absorbing of one cultural group into harmony with another)

s307.(the physical influence that produces a change in a physical quantity; "force equals mass times acceleration")

s308.(group of people willing to obey orders; "a public force is necessary to give security to the rights of citizens")

s309.(a group of people having the power of effective action; "he joined forces with a band of adventurers")

s310.(a unit that is part of some military service; "he sent Caesar a force of six thousand men")

s311.(a powerful effect or influence: "the force of his eloquence easily persuaded them")

s312.(physical energy or intensity: "he hit with all the force he could muster"; "it was destroyed by the strength of the gale"; "a government has not the vitality and forcefulness of a living man")

s313.((of a law) having legal validity; "the law is still in effect")

s314.(an act of aggression (as one against a person who resists); "he may accomplish by craft in the long run what he cannot do by force and violence in the short one")

s315.(a series of actions advancing a principle or tending toward a particular end; "he supported populist campaigns"; "they worked in the cause of world peace"; "the team was ready for a drive toward the pennant"; "the movement to end slavery"; "contributed to the war effort")

s316.(earnest and conscientious activity intended to do or accomplish something; "made an effort to cover all the reading material"; "wished him luck in his endeavor"; "she gave it a good try")

s317.(use of physical or mental energy; hard work; "he got an A for effort"; "they managed only with great exertion")

s318.(a notable achievement: "the book was her finest effort")

s319.(an operator that leaves unchanged the element on which it operates; "the identity under numerical multiplication is 1")

s320.(collective aspect of the set of characteristics by which a thing is recognizable or known)

s321.(exact sameness; "they shared an identity of interests")

s322.(the distinct personality of an individual regarded as a persisting entity: "you can lose your identity when you join the

The glossary of experiment 4

- army")
- s323**,(something that is lost or surrendered as a penalty;)
- s324**,(a penalty for a fault or mistake that involves losing or giving up something; "the contract specified forfeits if the work was not completed on time")
- s325**,(the act of losing or surrendering something as a penalty for a mistake or fault or failure to perform etc.)
- s326**,(the property possessed by something or someone of outstanding importance)
- s327**,(unusual largeness in size or extent)
- s328**,(the beginning or duration or completion or repetition of the action of a verb)
- s329**,(the visual percept of a region; "the most desirable feature of the park are the beautiful views")
- s330**,(a distinct feature or element in a problem; "he studied every facet of the question")
- s331**,(a characteristic to be considered)
- s332**,(the expression on a person's face; "a sad expression"; "a look of triumph"; "an angry face")
- s333**,(an organization of missionaries in a foreign land sent to carry on religious work)
- s334**,(a group of representatives or delegates)
- s335**,(an operation that is assigned by a higher headquarters; "the planes were on a bombing mission")
- s336**,(a task that has been assigned to a person or group; "a confidential mission to London"; "his charge was deliver a message")
- s337**,(the organized work of a religious missionary)
- s338**,(the contestant who wins the contest)
- s339**,(a gambler who wins a bet)
- s340**,(a person with a record of successes; "his son would never be the achiever that his father was"; "only winners need apply"; "if you want to be a success you have to dress like a success")
- s341**,(a unit of length equal to 4 inches; used in measuring horses; "the horse stood 20 hands")
- s342**,(a hired laborer on a farm or ranch; "the hired hand fixed the railing"; "a ranch hand")
- s343**,(a member of the crew of a ship; "all hands on deck")
- s344**,(a card player in a game of bridge; "we need a 4th hand for bridge")
- s345**,(a position given by its location to the side of an object; "objections were voiced on every hand")
- s346**,(the cards held in a card game by a given player at any given time; "I didn't hold a good hand all evening"; "he kept trying to see my hand")
- s347**,(a round of applause to signify approval; "give the little lady a great big hand")
- s348**,(something written by hand; "she recognized his handwriting"; "his hand was illegible")
- s349**,(one of two sides of an issue; "on the one hand..., but on the other hand...")
- s350**,(ability; "he wanted to try his hand at singing")
- s351**,(the (prehensile) extremity of the superior limb; "he had the hands of a surgeon"; "he extended his mitt")
- s352**,(a rotating pointer on the face of a timepiece; "the big hand counts the minutes")
- s353**,(terminal part of the forelimb in certain vertebrates (eg apes or kangaroos): "the kangaroo's forearms seem undeveloped but the powerful five-fingered hands are skilled at feinting and clouting"- Springfield (Mass.) Union)
- s354**,(physical assistance; "give me a hand with the chores")
- s355**,(a state of equilibrium)
- s356**,(equality of distribution)
- s357**,(an amount on the credit side of an account)
- s358**,(harmonious arrangement or relation of parts or elements within a whole (as in a design): "in all perfectly beautiful objects there is found the opposition of one part to another and a reciprocal balance"- John Ruskin)
- s359**,((mathematics) an attribute of a shape; exact correspondence of form on opposite sides of a dividing line or plane)
- s360**,(an equivalent counterbalancing weight)
- s361**,(a scale for weighing; depends on pull of gravity)
- s362**,(the psychological result of perception and learning and reasoning)
- s363**,(money in the form of bills or coins)
- s364**,(prompt payment for goods or services in currency or by check)
- s365**,(the situation in a contest in which the winner is undecided at the end; "the game ended in a draw"; "their record was 3 wins, 6 losses and a tie")
- s366**,(a social or business relationship: "a valuable financial affiliation"; "he was sorry he had to sever his ties with other members of the team"; "many close associations with England")
- s367**,((music) a slur over two notes of the same pitch; indicates that the note is to be sustained for their combined time value)
- s368**,(one of the cross braces that support the rails on a railway track; "the British call a railroad tie a sleeper")
- s369**,(a horizontal beam used to prevent two other structural members from spreading apart or separating; "he nailed the rafters together with a tie beam")
- s370**,(a cord (or string or ribbon or wire etc.) with which something is tied; "he needed a tie for the packages")
- s371**,(a long narrow piece of material worn (mostly by men) under a collar and tied in knot at the front; "he stood in front of the mirror tightening his necktie"; "he wore a vest and tie")
- s372**,(a fastener that serves to join or link; "the walls are held together with metal links placed in the wet mortar during construction")
- s373**,(an unstable situation of extreme danger or difficulty; "they went bankrupt during the economic crisis")
- s374**,(a crucial stage or turning point in the course of something; "after the crisis the patient either dies or gets better")
- s375**,(a member of an uncivilized people)
- s376**,(a crude uncouth ill-bred person lacking culture or refinement)
- s377**,(any entry into an area not previously occupied; "an invasion of tourists"; "an invasion of locusts"; "a viral invasion")
- s378**,(the act of invading; the act of an army that invades for conquest or plunder)
- s379**,(a sudden forceful flow)
- s380**,(a sudden or abrupt strong increase: "stimulated a surge of speculation"; "an upsurge of emotion"; "an upsurge in violent

The glossary of experiment 4

- crime")
- s381** (a state of agitation or turbulent change or development: "the political ferment produced a new leadership"; "social unrest")
- s382** (a feeling of restless agitation)
- s383** (the condition of things generally; "that's the way it is" or "I felt the same way")
- s384** (space for movement; "room to pass"; "make way for": "hardly enough elbow room to turn around")
- s385** (a portion of something divided into shares: "the split the loot three ways")
- s386** (a line leading to a place or point: "he looked the other direction"; "didn't know the way home")
- s387** (a general category of things; used in the expression "in the way of": "they didn't have much in the way of clothing")
- s388** (doing as one pleases or chooses: "if I had my way")
- s389** (the property of distance in general; "it's a long way to Moscow"; (colloquial) "he went a long way")
- s390** (a manner of performance; "a manner of living"; "in the characteristic New York style"; "a way of life")
- s391** (any road or path affording passage from one place to another; "he said he was looking for the way out")
- s392** (a journey or passage; "they are on the way")
- s393** (how a result is obtained or an end is achieved; "a means of communication"; "an example is the best agency of instruction"; "the true way to success")
- s394** (the interval between two times; "the distance from birth to death"; "it all happened in the space of 10 minutes")
- s395** (an empty area (usually bounded in some way between things); "the architect left space in front of the building"; "they stopped at an open space in the jungle"; "the space between his teeth")
- s396** (an area reserved for some particular purpose; "the laboratory's floor space")
- s397** (any region in space outside the Earth's atmosphere; "the astronauts walked in space without a tether")
- s398** ((mathematics) any set of points that satisfy a set of postulates of some kind; "assume the vector space is finite dimensional")
- s399** (one of the areas between or below or above the lines of a musical staff; "the spaces are the notes F-A-C-E")
- s400** (a blank character used to separate successive words in writing or printing; "he said the space is the most important character in the alphabet")
- s401** (a blank area; "write your name in the space provided")
- s402** ((printing) a block of type without a raised letter; used for spacing between words)
- s403** (the unlimited 3-dimensional expanse in which everything is located; "they tested his ability to locate objects in space")
- s404** (the flow of air that is driven backwards by an aircraft propeller)
- s405** ((biology) a taxonomic group that is a division of a species; usually arises as a consequence of geographical isolation within a species)
- s406** (people who are believed to belong to the same genetic stock; "some biologists doubt that there are important genetic differences between races of human beings")
- s407** (any competition; "the race for the presidency")
- s408** (a contest of speed; "the race is to the swift")
- s409** (a canal for a current of water)
- s410** (a search for knowledge; "their pottery deserves more research than it has received")
- s411** (systematic investigation to establish facts)
- s412** (buildings with facilities for manufacturing)
- s413** (any of various water-soluble compounds capable of turning litmus blue and reacting with an acid to form a salt and water; "bases include oxides and hydroxides of metals and ammonia")
- s414** (the bottom side of a geometric figure from which the altitude can be constructed; "the base of the triangle")
- s415** ((in a digital numeration system) the positive integer that is equivalent to one in the next higher counting place; "10 is the radix of the decimal system")
- s416** (the bottom or lowest part; "the base of the mountain")
- s417** ((anatomy) the part of an organ nearest its point of attachment: "the base of the skull")
- s418** (the place where you are stationed and from which missions start and end)
- s419** (a lower limit: "the government established a wage floor")
- s420** ((linguistics) the form of a word after all affixes are removed; "thematic vowels are part of the stem")
- s421** (the fundamental assumptions underlying an explanation; "the whole argument rested on a basis of conjecture")
- s422** (the basic facilities and equipment needed for the functioning of a country or area; "the industrial base of Japan")
- s423** (lowest supporting part of a structure; "it was built on a base of solid rock"; "he stood at the foot of the tower")
- s424** (the principal ingredient of a mixture; "glycerinated gelatin is used as a base for many ointments"; "he told the painter that he wanted a yellow base with just a hint of green"; "everything she cooked seemed to have rice as the base")
- s425** (a flat bottom on which something is intended to sit; "a tub should sit on its own base")
- s426** (installation from which a military force initiates operations; "the attack wiped out our forward bases")
- s427** ((electronics) the part of a transistor that separates the emitter from the collector)
- s428** (place that runner must touch before scoring; "he scrambled to get back to the bag")
- s429** (a support or foundation; "the base of the lamp")
- s430** (a vehicle capable of traveling in outer space; technically a satellite around the sun)
- s431** (something determined in relation to something that includes it; "he wanted to feel a part of something bigger than himself"; "I read a portion of the manuscript"; "the smaller component is hard to reach")
- s432** (an abstract part of something; "jealousy was a component of his character"; "two constituents of a musical composition are melody and harmony"; "the grammatical elements of a sentence"; "a key factor in her success"; "humor: an effective ingredient of a speech")
- s433** (an artifact that is one of the individual parts of which a composite entity is made up; especially a part that can be separated from or attached to a system: "spare components for cars"; "a component or constituent element of a system")
- s434** (the branch of military science dealing with military command and the planning and conduct of a war)
- s435** (an elaborate and systematic plan of action)

The glossary of experiment 4

- s436**, (the age at which a person is considered competent to manage their own affairs)
- s437**, ((in an election) more than half of the votes)
- s438**, (the property resulting from being or relating to the greater in number of two parts; the main part; "the majority of his customers prefer it"; "the bulk of the work is finished")
- s439**, (any object with a spherical shape; "a ball of fire")
- s440**, (the 3rd planet from the sun; the planet on which we live; "the Earth moves around the sun"; "he sailed around the world")
- s441**, (a sphere on which a map (especially of the earth) is represented)
- s442**, (a state of prosperity or fame; "he is enjoying great success"; "he does not consider wealth synonymous with success")
- s443**, (an event that accomplishes its intended purpose; "let's call heads a success and tails a failure"; "the election was a remarkable success for Republicans")
- s444**, (an attainment that is successful; "his success in the marathon was unexpected"; "his new play was a great success")
- s445**, (someone who tries to bring peace)
- s446**, (the territory occupied by a nation; "he returned to the land of his birth"; "he visited several European countries")
- s447**, (a federation of tribes (especially native American tribes); "the Shawnee nation")
- s448**, (a politically organized body of people under a single government; "the state has elected a new president")
- s449**, (the people of a nation or country or a community of persons bound by a common heritage; "a nation of Catholics"; "the whole country worshipped him")
- s450**, (a demanding or stimulating situation; "they reacted irrationally to the challenge of Russian power")
- s451**, (a call to engage in a contest or fight)
- s452**, (a formal objection to the selection of a particular person as a juror)
- s453**, (questioning a statement and demanding an explanation; "he challenged the assumption that Japan is our enemy")
- s454**, (a demand by a sentry for a password or identification)
- s455**, (the act of accomplishing something)
- s456**, (any piece of work)
- s457**, (a specific piece of work required to be done as a duty or for a specific fee: "estimates of the city's loss on that job ranged as high as a million dollars"; "the job of repairing the engine took several hours"; "the endless task of classifying the samples"; "the farmer's morning chores")
- s458**, (general line of orientation: "the river takes a southern course"; "the northeastern trend of the coast")
- s459**, (a connected series of events or actions or developments; "the government took a firm course" or "historians can only point out those lines for which evidence is available")
- s460**, (part of a meal served at one time; "she prepared a three course meal")
- s461**, (a layer of masonry; "a course of bricks")
- s462**, (a circumscribed area of land or water laid out for a sport; "the course had only nine holes"; "the course was less than a mile")
- s463**, (education imparted in a series of lessons or class meetings; "he took a course in basket weaving"; "flirting is not unknown in college classes")
- s464**, (a mode of action; "if you persist in that course you will surely fail")
- s465**, (a successful ending of a struggle or contest; "the general always gets credit for his army's victory"; "the agreement was a triumph for common sense")
- s466**, (an amount due (as at a restaurant or bar); "add it to my score and I'll settle later")
- s467**, (a notch that is made to keep a tally)
- s468**, (a number that expresses the accomplishment of a team or an individual in a game or contest; "the score was 7 to 0")
- s469**, (grounds; "don't do it on my account"; "the paper was rejected on account of its length"; "he tried to blame the victim but his success on that score was doubtful")
- s470**, (a set of twenty members; "four score and seven years ago")
- s471**, (a resentment strong enough to justify retaliation; "holding a grudge"; "settling a score")
- s472**, (a written form of a musical composition; parts for different instruments appear on separate staves on large pages; "he studied the score of the sonata")
- s473**, (the facts about an actual situation; "he didn't know the score")
- s474**, (a number or letter indicating quality (especially of a student's performance); "she made good marks in algebra"; "grade A milk"; "what was your score on your homework?")
- s475**, (the act of scoring in a game or sport; "the winning score came with less than a minute left to play")
- s476**, (a seduction culminating in sexual intercourse; "calling his seduction of the girl a 'score' was a typical example of male slang")
- s477**, (a person who backs a politician or a team etc.; "all their supporters came out for the game"; "they are friends of the library")
- s478**, (a person you know well and regard with affection and trust; "he was my best friend at the university")
- s479**, (an associate who provides assistance; "he's a good ally in fight"; "they were friends of the workers")
- s480**, (a person with whom you are acquainted; "I have trouble remembering the names of all my acquaintances"; "we are friends of the family")
- s481**, (the use of closed-class words instead of inflections: eg, "the father of the bride" instead of "the bride's father")
- s482**, (a form of literary criticism in which the structure of a piece of writing is analyzed)
- s483**, (a branch of mathematics involving calculus and the theory of limits; sequences and series and integration and differentiation)
- s484**, (the abstract separation of a whole into its constituent parts for study)
- s485**, (a set of techniques for exploring underlying motives and a method of treating various mental disorders; "his physician recommended psychoanalysis")
- s486**, (an investigation of the component parts of a whole)
- s487**, (confirmation that some fact or statement is true)
- s488**, (program listings or technical manuals describing the operation and use of programs)

The glossary of experiment 4

- s489**,(documentary validation; "his documentation of the results was excellent"; "the strongest support for this this view is the work of Jones")
- s490**,(people descended from a common ancestor; "his family had lived in Massachusetts since the Mayflower")
- s491**,(people in general; "they're just country folk"; "the common people determine the group character and preserve its customs from one generation to the next")
- s492**,(the traditional and typically anonymous music that is an expression of the life of people in a community)
- s493**,(an expert at calculation (or at operating calculating machines))
- s494**,(a machine for performing calculations automatically)
- s495**,(the side of an object that is opposite its front; "his room was toward the rear of the hotel")
- s496**,(the part of something that is furthest from the normal viewer: "he stood at the back of the stage"; "it was hidden in the rear of the store")
- s497**,(the back of a military formation or procession; "infantrymen were in the rear")
- s498**,(the fleshy part of the human body that you sit on)
- s499**,(the side that goes last or is not normally seen; "he wrote the date on the back of the photograph")
- s500**,(the tip of an abscess (where the pus accumulates))
- s501**,(the length or height based on the size of a human or animal head; "he is two heads taller than his little sister"; "his horse won by a head")
- s502**,(a dense clusters of flowers or foliage: "a head of cauliflower"; "a head of lettuce")
- s503**,(the pressure exerted by a fluid; "a head of steam")
- s504**,(the educator who has executive authority for a school; "she sent unruly pupils to see the principal")
- s505**,(an individual person; "tickets are \$5 per head")
- s506**,(a person who is in charge; "the head of the whole operation")
- s507**,((informal) a user of (usually soft) drugs; "the office was full of secret heads")
- s508**,(the foam or froth that accumulates at the top when you pour an effervescent liquid into a container; "the beer had a large head of foam")
- s509**,(a rounded compact mass; "the head of a comet")
- s510**,(the top of something; "the head of the stairs"; "the head of the page"; "the head of the list")
- s511**,(the part in the front or nearest the viewer; "he was in the forefront"; "he was at the head of the column")
- s512**,(the source of water from which a stream arises; "they tracked him back toward the head of the stream")
- s513**,(the front of a military formation or procession; "the head of the column advanced boldly"; "they were at the head of the attack")
- s514**,(a difficult juncture; "a pretty pass"; "matters came to a head yesterday")
- s515**,(forward movement; "the ship made little headway against the gale")
- s516**,(a V-shaped mark at one end of an arrow pointer; "the point of the arrow was due north")
- s517**,(the subject matter at issue; "the question of disease merits serious discussion"; "under the head of minor Roman poets")
- s518**,(a line of text serving to indicate what the passage below it is about; "the heading seemed to have little to do with the text")
- s519**,((linguistics) the word in a constituent that plays the same grammatical role as the whole)
- s520**,(that which is responsible for one's thoughts and feelings; the seat of the faculty of reason; "his mind wandered"; "I couldn't get his words out of my head")
- s521**,(the upper or front part of the body in animals; contains the face and brains; "he stuck his head out the window")
- s522**,((computer science) a tiny electromagnetic coil and metal pole used to write and read magnetic patterns on a disk)
- s523**,((usually plural) an obverse side of a coin that bears the representation of a person's head; "call heads or tails!")
- s524**,(the striking part of a tool; "the head of the hammer")
- s525**,(a toilet on board a boat of ship)
- s526**,(a part that projects out from the rest; "the head of the nail"; "a pinhead is the head of a pin")
- s527**,(a membrane that is stretched taut over a drum)
- s528**,(a single domestic animal: "200 head of cattle")
- s529**,(oral-genital stimulation; "they say he gives good head")
- s530**,(a grammatical category of pronouns and verb forms; "stop talking about yourself in the third person")
- s531**,(a person's body (usually including their clothing); "a weapon was hidden on his person")
- s532**,(a human being; "there was too much for one person to do")
- s533**,((informal) uncomplimentary terms for a policeman)
- s534**,(an investor with an optimistic market outlook)
- s535**,(a very large and strong man; "he was a bull of a man")
- s536**,(the center of a target)
- s537**,(a ludicrously false statement)
- s538**,(uncastrated adult male of domestic cattle)
- s539**,(mature male of various mammals of which the female is called 'cow'; eg whales or elephants or especially cattle)
- s540**,(a serious and ludicrous blunder; "he made a bad bull of the assignment")
- s541**,((sports) a preliminary competition to determine qualifications; "the trials for the semifinals began yesterday")
- s542**,(an annoying or frustrating event; "his mother-in-law's visits were a great trial for him"; "life is full of tribulations"; "a visitation of the plague")
- s543**,(trying something to find out about it; "a sample for ten days free trial"; "a trial of progesterone failed to relieve the pain")
- s544**,((law) the determination of a person's innocence or guilt by due process of law; "he had a fair trial and the jury found him guilty")
- s545**,((law) legal proceedings consisting of the judicial examination of issues by a competent tribunal; "most of these complaints are settled before they go to trial")
- s546**,(the act of undergoing testing; "he survived the great test of battle"; "candidates must compete in a trial of skill")
- s547**,(the act of testing something; "in the experimental trials the amount of carbon was measured separately"; "he called each

The glossary of experiment 4

flip of the coin a new trial")

s548.(part of a statement that is not correct; "the book was full of errors")

s549.(a misconception resulting from incorrect information)

s550.(departure from what is ethically acceptable)

s551.(inadvertent incorrectness)

s552.((baseball) a failure of a defensive player to make an out when normal play would have sufficed)

s553.(a wrong action attributable to bad judgment or ignorance or inattention; "the fault was all mine")

s554.(verbal wit (often at another's expense but not to be taken seriously); "he became a figure of fun")

s555.(a disposition to find (or make) causes for amusement; "her playfulness surprised me"; "he was fun to be with")

s556.(violent and excited activity; "she asked for money and then the fun began"; (colloquial) "they began to fight like fun")

s557.(activities that are enjoyable or amusing; "I do it for the fun of it"; "he is fun to have around")

s558.(a state of difficulty that needs to be resolved; "she and her husband are having problems"; "it is always a job to contact him"; "urban problems such as traffic congestion and smog")

s559.(a question raised for consideration or solution; "our homework consisted of ten problems to solve")

s560.(a source of difficulty: "one trouble after another delayed the job"; "what's the problem?")

s561.((biology) the taxonomic group whose characteristics are used to define the next higher taxon)

s562.(printed characters; "small type is hard to read")

s563.(all of the tokens of the same symbol; "the word 'element' contains five different types of character")

s564.(a subdivision of a particular kind of thing; "what type of sculpture do you prefer?")

s565.(a small block of metal bearing a raised character on one end; produces a printed character when inked and pressed on paper; "he dropped a case of type so they made him pick them up")

s566.(a person who participates in or is skilled at some game)

s567.(someone who plays a musical instrument (as a profession))

s568.(a theatrical performer)

s569.(proper or appropriate position or location; "a woman's place is no longer in the kitchen")

s570.(proper or designated social situation: "he overstepped his place"; "the responsibilities of a man in his station"; "married above her station")

s571.(a particular situation: "If you were in my place what would you do?")

s572.(a point located with respect to surface features of some region; "this is a nice place for a picnic")

s573.(a space reserved for sitting (as in a theater or on a train or airplane); "he booked their seats in advance"; "he sat in someone else's place")

s574.(a general vicinity; "He comes from a place near Chicago")

s575.(the particular portion of space occupied by a physical object: "he put the lamp back in its place")

s576.(a public square with room for pedestrians; "they met at Elm Plaza"; "Grosvenor Place")

s577.(where you live; "deliver the package to my home"; "he doesn't have a home to go to"; "your place or mine?")

s578.(any area set aside for a particular purpose; "who owns this place?")

s579.((in horse racing) a finish in second place)

s580.(an item on a list or in a sequence; "in the second place"; "moved from third to fifth position")

s581.(the passage that is being read; "he lost his place on the page")

s582.(an abstract mental location; "he has a special place in my thoughts"; "a place in my heart"; "a political system with no place for the less prominent groups")

s583.(the function or position properly or customarily occupied or served by another: "can you go in my stead?"; "took his place"; "in lieu of")

s584.(a job in an organization or hierarchy; "he occupied a post in the treasury")

s585.(a pause for relaxation; "people actually accomplish more when they take time for short rests")

s586.(a state of inaction; "a body will continue in a state of rest until acted upon")

s587.(euphemisms for death (based on an analogy between lying in a bed and in a tomb); "she was laid to rest beside her husband"; "they had to put their family pet to sleep")

s588.(something left after other parts have been taken away; "there was no remainder"; "he threw away the rest")

s589.(a musical notation indicating a silence of a specified duration)

s590.(a support on which things can be put; "the gun was steadied on a special rest")

s591.(freedom from activity (work or strain or responsibility); "took his repose by the swimming pool")

s592.(a distinct period in history or in a person's life; "the industrial revolution opened a new chapter in British history"; "the divorce was an ugly chapter in their relationship")

s593.(an ecclesiastical assembly of the monks in a monastery or even of the canons of a church)

s594.(a local branch of some fraternity or association; "he joined the Atlanta chapter")

s595.(a series of related events forming an episode; "a chapter of disasters")

s596.(a subdivision of a written work; usually numbered and titled; "he read a chapter every night before falling asleep")

s597.(a condition of regular or proper arrangement: "he put his desk in order"; "put the chessmen in order")

s598.(established customary state especially of society; "order ruled in the streets"; "law and order")

s599.(logical or comprehensible arrangement of separate elements of a group; "we shall consider these questions in the inverse order of their presentation")

s600.((biology) taxonomic group containing one or more families)

s601.((often plural) a command given by a superior (eg, a military or law enforcement officer) that must be obeyed; "the British ships dropped anchor and waited for orders from London")

s602.(a body of rules followed by an assembly)

s603.(a legally binding command or decision entered on the court record (as if issued by a court or judge); "a friend in New Mexico said that the order caused no trouble out there")

s604.(a commercial document used to request someone to supply something in return for payment; "IBM received an order for a hundred computers")

The glossary of experiment 4

- s605**, (a degree in a continuum of size or quantity; "it was on the order of a mile"; "an explosion of a low order of magnitude")
- s606**, (putting in order; "there were mistakes in the ordering of items on the list")
- s607**, ((law) a courtroom conference between the lawyers and the judge in a trial that is held out of the jury's hearing)
- s608**, (a short news story presenting sidelights on a major story)
- s609**, (the lowest tone of a harmonic series)
- s610**, (a concluding summary (as in presenting a case before a law court))
- s611**, (a relational difference between states; especially between states before and after some event: "he attributed the change to their marriage")
- s612**, (coins of small denomination regarded collectively; "he had a pocketful of change")
- s613**, (the balance of money received when the amount you tender is greater than the amount due; "I paid with a twenty and pocketed the change")
- s614**, (money received in return for its equivalent in a larger denomination or a different currency; "he got change for a twenty and used it to pay the taxi driver")
- s615**, (the result of alteration or modification; "there were marked changes in the lining of the lungs"; "there had been no change in the mountains")
- s616**, (an event that occurs when something passes from one state or phase to another: "the change was intended to increase sales"; "this storm is certainly a change for the worse")
- s617**, (a thing that is different; "he inspected several changes before selecting one")
- s618**, (a different or fresh set of clothes; "she brought a change in her overnight bag")
- s619**, (the act of changing something; "the change of government had no impact on the economy"; "his change on abortion cost him the election")
- s620**, (a special situation; "this thing has got to end"; "it is a remarkable thing")
- s621**, (a persistent illogical feeling of desire or aversion: "he has a thing about seafood"; "she has a thing about him")
- s622**, (an event: "a funny thing happened on the way to the...")
- s623**, (a statement regarded as an object; "to say the same thing in other terms" or "how can you say such a thing?")
- s624**, (a special objective: "the thing is to stay in bounds")
- s625**, (a special abstraction; "a thing of the spirit"; "things of the heart")
- s626**, (a vaguely specified concern; "several matters to attend to"; "it is none of your affair"; "things are going well")
- s627**, (any attribute or quality considered as having its own existence: "the thing I like about her is ...")
- s628**, (an entity that is not named specifically; "I couldn't tell what the thing was")
- s629**, (an artifact; "how does this thing work?")
- s630**, (an action; "how could you do such a thing?")
- s631**, (reproduction by applying ink to paper as for publication)
- s632**, (all the copies of a work printed at one time; "they ran off an initial printing of 2000 copies")
- s633**, (text written in the style of printed matter)
- s634**, (the business of printing)
- s635**, (the capacity to attract and hold something)
- s636**, (strip sewn over or along an edge for reinforcement or decoration)
- s637**, (the front and back covering of a book; "the book had a leather binding")
- s638**, (the act of applying a bandage)
- s639**, (a squeeze with the fingers)
- s640**, (an inherited pattern of thought or action)
- s641**, (a specific practice of long standing)
- s642**, (a publication containing a variety of works)
- s643**, (a concise but comprehensive summary of a larger work)
- s644**, ((computer science) a graphic symbol (usually a simple picture) that denotes a program or a command or a data file or a concept in a graphical user interface)
- s645**, (a visual representation of an object or scene or person produced on a surface; "they showed us the pictures of their wedding"; "a movie is a series of images projected so rapidly that the eye integrates them")
- s646**, (a conventional religious picture painted in oil on a small wooden panel; venerated in the Eastern Church)
- s647**, (a record in which commercial accounts are recorded; "they got a subpoena to examine our books")
- s648**, (a written version of a play or other dramatic composition; used in preparing for a performance)
- s649**, (a compilation of the known facts regarding something or someone; "Al Smith used to say, 'Let's look at the record'; "his name is in all the recordbooks")
- s650**, (a copy of a written work or composition that has been published (printed on pages bound together); "I am reading a good book on economics")
- s651**, (a major division of a long written composition; "the book of Isaiah")
- s652**, (an accounting book as a physical object: "he bought a new daybook")
- s653**, (a number sheets (ticket or stamps etc.) bound together on one edge; "he bought a book of stamps")
- s654**, (a book as a physical object: a number of pages bound together; "he used a large book as a doorstop")
- s655**, (an expert who gives advice)
- s656**, (a passage from the Bible that is used as the subject of a sermon; "the preacher chose a text from Psalms to introduce his sermon")
- s657**, (a book prepared for use in schools or colleges; "his economics textbook is in its tenth edition")
- s658**, (the main body of a written work (as distinct from illustrations or footnotes etc.); "pictures made the text easier to understand")
- s659**, (the words of something written; "there were more than a thousand words of text"; "they handed out the printed text of the mayor's speech"; "he wants to reconstruct the original text")
- s660**, (a publication (or a passage from a publication) that is referred to; "he carried an armful of references back to his desk"; "he spent hours looking for the source of that quotation")

The glossary of experiment 4

- s661**, (an indicator that orients you generally; "it is used as a reference for comparing the heating and the electrical energy involved")
- s662**, (a remark that calls attention to something or someone; "she made frequent mention of her promotion"; "there was no mention of it"; "the speaker made several references to his wife")
- s663**, (a short note acknowledging a source of information or quoting a passage; "the student's essay failed to list several important citations"; "the article includes mention of similar clinical cases")
- s664**, (a formal recommendation by a former employer to a potential future employer describing the person's qualifications and dependability; "requests for character references are all to often answered evasively")
- s665**, (a book to which you can refer for authoritative facts; "he contributed articles to the basic reference work on that topic")
- s666**, (the relation between a word or phrase and the object or idea it refers to; "he argued that reference is a consequence of conditioned reflexes")
- s667**, (the class of objects that an expression refers to; "the extension of 'satellite of Mars' is the set containing only Demos and Phobos")
- s668**, (the act of referring; "reference to an encyclopedia produced the answer")
- s669**, (an individual or group or structure or other entity regarded as a structural or functional constituent of a whole; "the reduced number of units and installations"; "the word is a basic linguistic unit")
- s670**, (any division of quantity accepted as a standard of measurement or exchange; "the dollar is the United States unit of currency"; "a unit of wheat is a bushel"; "change per unit volume")
- s671**, (a single undivided natural entity occurring in the composition of something else; "units of nucleic acids")
- s672**, (an organization regarded as part of a larger social group; "the coach said the offensive unit did a good job"; "after the battle the soldier had trouble rejoining his unit")
- s673**, (a single undivided whole; "an idea is not a unit that can be moved from one brain to another")
- s674**, (an assemblage of parts that is regarded as a single entity; "how big is that part compared to the whole?"; "the repairman simply replaced the unit")
- s675**, (a combination of interrelated interacting artifacts designed to work as a coherent entity; "he bought a new stereo system"; "the unit consists of a motor and a small computer")
- s676**, (a condition superior to an earlier condition; "the new school represents a great improvement")
- s677**, (a change for the better; progress in development)
- s678**, (the act of improving something; "Their improvements increased the value of the property")
- s679**, (the organization that is the governing authority of a political unit; "the government reduced taxes"; "the matter was referred to higher authorities")
- s680**, (the study of government of states and other political units)
- s681**, (the system or form by which a community or other political unit is governed; "tyrannical government")
- s682**, (the act of governing; exercising authority; "regulations for the government of state prisons"; "he had considerable experience of government")
- s683**, (a piece of ground having specific characteristics or military potential; "they decided to attack across the rocky terrain")
- s684**, (the score needed to win a game; "he is serving for the game")
- s685**, (the flesh of wild animals that is used for food)
- s686**, (the equipment needed to play a game; "the child received several games for his birthday")
- s687**, (animal hunted for food or sport)
- s688**, (informal terms for your occupation; "he's in the plumbing game"; "she's in show biz")
- s689**, (an amusement or pastime; "he thought of his painting as a game that filled his empty time"; "his life was all fun and games")
- s690**, (a single play of a game; "the game lasted 2 hours")
- s691**, (a contest with rules to determine a winner; "you need four people to play this game")
- s692**, (an abstract or general idea inferred or derived from specific instances)
- s693**, (a society in an advanced state of development)
- s694**, (everything that exists anywhere; "they study the evolution of the universe"; "the biggest tree in existence")
- s695**, (a part of the earth that can be considered separately; "the outdoor world"; "the world of insects")
- s696**, (people in general considered as a whole; "he is a hero in the eyes of the public")
- s697**, (people in general; especially a distinctive group of people with some shared interest; "the Western world")
- s698**, (all of the inhabitants of the earth; "all the world loves a lover")
- s699**, (all of your experiences that determine how things appear to you; "his world was shattered"; "we live in different worlds"; "for them demons were as much a part of reality as trees were")
- s700**, (the concerns of the world as distinguished from heaven and the afterlife; "they consider the church to be independent of the world")
- s701**, (a written record of a commercial transaction)
- s702**, (a message submitted in a competition)
- s703**, (an item inserted in a written record)
- s704**, (something that allows entry or exit; "they waited at the entrance to the garden"; "beggars waited just outside the entryway to the cathedral")
- s705**, (the act of beginning something new; "they looked forward to the debut of their new product line")
- s706**, (the act of entering; "she made a grand entrance")
- s707**, (a freedom from financial difficulty that promotes a comfortable state; "a life of luxury and ease"; "he had all the material comforts of this world")
- s708**, (the condition of being comfortable or relieved (especially after being relieved of distress); "he enjoyed his relief from responsibility"; "getting it off his conscience gave him some ease")
- s709**, (freedom from constraint or embarrassment; "I am never at ease with strangers")
- s710**, (freedom from difficulty or hardship or effort; "he rose through the ranks with apparent ease"; "they put it into containers for ease of transportation")

A-5 Master results from experiments

Table A-5.1 Master scoring table all experiments

Sequence	Instances	lexemes	Decided (I)	Decided (L)	Vertices	Edges	Comps	Integrity	Score	Sum of scores of isolated p's	Score pr Instance	Score pr Lexeme	Average of Sum of scores of individual paragraphs		Relative gain from joined paragraphs	
													pr instance	pr lexeme	pr instance	pr lexeme
dall-p0	19	14	0,63	0,50	7	4	3	19%	15	-	0,79	1,07	-	-	-	-
dall-p1	23	20	0,09	0,10	2	1	1	100%	2	-	0,09	0,10	-	-	-	-
dall-p2	29	25	0,17	0,16	4	2	2	33%	5	-	0,17	0,20	-	-	-	-
dall-p3	19	18	0,00	0,00	0	0	0	0%	0	-	0,00	0,00	-	-	-	-
dall-p4	18	16	0,00	0,00	0	0	0	0%	0	-	0,00	0,00	-	-	-	-
dall-p5	17	15	0,53	0,47	7	4	3	19%	11	-	0,65	0,73	-	-	-	-
dall-p6	19	19	0,00	0,00	0	0	0	0%	0	-	0,00	0,00	-	-	-	-
dall-p7	27	26	0,15	0,15	4	2	2	33.3%	4	-	0,15	0,15	-	-	-	-
dall-p8	60	44	0,27	0,27	11	5	6	9%	16	-	0,27	0,36	-	-	-	-
d0X-0	42	32	0,36	0,28	9	5	4	14%	18	17	0,43	0,56	0,40	0,53	0,02	0,03
d0X-1	48	38	0,40	0,32	12	7	5	11%	24	20	0,50	0,63	0,42	0,53	0,08	0,11
d0X-2	38	32	0,32	0,22	7	4	3	19%	15	15	0,39	0,47	0,39	0,47	0,00	0,00
d0X-3	37	28	0,35	0,25	7	4	3	20%	18	15	0,49	0,64	0,41	0,54	0,08	0,11
d0X-4	36	25	0,69	0,56	14	8	6	9%	31	26	0,86	1,24	0,72	1,04	0,14	0,20
d0X-5	35	29	0,37	0,24	7	4	3	19%	17	15	0,49	0,59	0,43	0,52	0,06	0,07
d0X-6	46	38	0,35	0,26	10	6	4	13%	21	19	0,46	0,55	0,41	0,50	0,04	0,05
d0X-7	79	56	0,35	0,30	16	9	7	8%	37	31	0,47	0,66	0,39	0,55	0,08	0,11
d0-5_6_7_8-0	125	93	39,00	33,00	31	18	13	4%	61	33	0,49	0,66	0,26	0,35	0,22	0,30
d0-5_6_7_8-1	19	19	0,00	0,00	0	0	0	0%	0	0	0,00	0,00	0,00	0,00		
d0-5_6_7_8-2	27	26	0,15	0,15	4	2	2	33%	4	4	0,15	0,15	0,15	0,15		
d0-5_6_7_8-3	60	44	0,27	0,27	11	5	6	9%	16	16	0,27	0,36	0,27	0,36		
dall-0	228	160	0,43	0,39	61	38	23	2%	116	89	0,51	0,73	0,39	0,56	0,12	0,17

Table A-5.2 Experiment 1 - isolated paragraphs

	dall-p0	dall-p1	dall-p2	dall-p2	dall-p2	dall-p3	dall-p4	dall-p5	dall-p6	dall-p7	dall-p7	dall-p7	dall-p8	dall-p8	dall-p8
Interpretation	1	1	1	2	3	1	1	1	1	1	2	3	1	2	3
Score	15	2	5	4	4	0	0	11	0	4	4	4	16	15	15
token															
achievement								s455							
aggression															
area		s57													
arts								s267							
aspect															
binding															s637
book													s650	s651	s654
case															
cash															
challenge															
change													s617	s617	s617
chapter													s596	s594	s594
civilization	s693							s693							
component															
conquest															
course										s458	s464	s464			
culture	s292							s292							
darkness															
defense															
demand															
discovery															
effort								s318							
empire															
feature															
folk															
force	s312														
frequency															
friend															
game												s684			
goal	s4														
government															
greatness															
hand															
head													s517	s517	s517
history								s287							
importance															
impulse	s19														
knowledge															
measure															
nation															
objective	s14														
opponent															

order													s299	s299
path														
people														
person														
place												s575	s575	s575
population														
power		s145	s145	s145										
problem												s559	s559	s559
production														
purpose														
range														
rate		s107	s108	s105										
rear												s496	s496	s496
reference												s665		
rest														
science		s181	s181	s181										
score								s475	s475	s468				
society	s300						s300							
space														
speed			s114	s112										
sucess								s444	s444					
taste														
tax		s178												
territory	s44													
text												s659	s659	
thing												s628	s628	s628
trial														
tribe														
type														
unit														
value														
variety														
way								s386	s195	s195				
wiiner														
world														

Table A-5.3 Experiment 2 - first and second paragraph

	dall-p0	dall-p1	d0X-0	d0X-0		dall-p0	dall-p1	d0X-0	d0X-0
Interpretation	#1	#1	#1	#2	Interpretation	#1	#1	#1	#2
Score	15	2	18	18	Score	15	2	18	18
token					token				
achievement	-----	-----			nation	-----	-----		
aggression	-----	-----			objective	s14		s14	s14
area	-----	s57	s57	s57	opponent	-----	-----		
arts	-----	-----			order	-----	-----		
aspect	-----	-----			path	-----	-----		
binding	-----	-----			people	-----	-----		
book	-----	-----			person	-----	-----		
case	-----	-----			place	-----	-----		
cash	-----	-----			population	-----	-----		
challenge	-----	-----			power	-----	-----		
change	-----	-----			problem	-----	-----		
chapter	-----	-----			production	-----	-----		
civilization	s693	-----	s693	s693	purpose	-----	-----		
component	-----	-----			range	-----	-----		
conquest	-----	-----			rate	-----	-----		
course	-----	-----			rear	-----	-----		
culture	s292	-----	s292	s292	reference	-----	-----		
darkness	-----	-----			rest	-----	-----		
defense	-----	-----			science	-----	-----		
demand	-----	-----			score	-----	-----		
discovery	-----	-----			society	s300		s300	s300
effort	-----	-----			space	-----	-----		
empire	-----	-----			speed	-----	-----		
feature	-----	-----			success	-----	-----		
folk	-----	-----			taste	-----	-----		
force	s312	-----	s312	s310	tax	-----	-----		
frequency	-----	-----			territory	-----	s44	s44	s44
friend	-----	-----			text	-----	-----		
game	-----	-----			thing	-----	-----		
goal	s4	-----	s4	s4	trial	-----	-----		
government	-----	-----			tribe	-----	-----		
greatness	-----	-----			type	-----	-----		
hand	-----	-----			unit	-----	-----		s672
head	-----	-----			value	-----	-----		
history	-----	-----			variety	-----	-----		
importance	-----	-----			way	-----	-----		
impulse	s19	-----	s19	-----	winner	-----	-----		
knowledge	-----	-----			world	-----	-----		
measure	-----	-----							

Table A-5.4 Experiment 2 - first and third paragraph

	dall-p0	dall-p2	dall-p2	d0X-1	d0X-1	d0X-1		dall-p0	dall-p2	dall-p2	d0X-1	d0X-1	d0X-1
Interpretation	#1	#1	#2	#1	#2	#3	Interpretation	#1	#1	#2	#1	#2	#3
Score	15	5	4	24	23	23	Score	15	5	4	24	23	23
token							token						
achievement							objective	s14			s14	s14	s14
aggression							opponent						
area							order						
arts							path						
aspect							people						
binding							person						
book							place						
case							population						
cash							power	s145	s145		s145	s145	s145
challenge							problem						
change							production						
chapter							purpose				s109	s109	s109
civilization	s693			s693	s693	s693	range						
component							rate	s107	s108		s107	s108	s105
conquest							rear						
course							reference						
culture	s292			s292	s292	s292	rest						
darkness							science	s181	s181		s181	s181	s181
defense							score						
demand							society	s300			s300	s300	s300
discovery							space						
effort							speed		s114		s114	s112	
empire							success						
feature							taste						
folk							tax	s178			s178		
force	s312			s312	s312	s312	territory						
frequency							text						
friend							thing						
game							trial						
goal	s4			s4	s4	s4	tribe						
government							type						
greatness							unit						
hand							value						
head							variety						
history							way						
importance							winner						
impulse	s19			s19	s19	s19	world						
knowledge													
measure													
nation													

Table A-5.5 Experiment 2 - first and fourth paragraph

	dall-p0	dall-p3	d0X-2	d0X-2		dall-p0	dall-p3	d0X-2	d0X-2
Interpretation	#1	#1	#1	#2	Interpretation	#1	#1	#1	#2
Score	15	0	15	14	Score	15	0	15	14
token					token				
achievement	-----	-----			opponent	-----	-----		
aggression	-----	-----			order	-----	-----		
area	-----	-----			path	-----	-----		
arts	-----	-----			people	-----	-----		
aspect	-----	-----			person	-----	-----		
binding	-----	-----			place	-----	-----		
book	-----	-----			population	-----	-----		
case	-----	-----			power	-----	-----		
cash	-----	-----			problem	-----	-----		
challenge	-----	-----			production	-----	-----		
change	-----	-----			purpose	-----	-----		
chapter	-----	-----			range	-----	-----		
civilization	s693		s693	s693	rate	-----	-----		
component	-----	-----			rear	-----	-----		
conquest	-----	-----			reference	-----	-----		
course	-----	-----			rest	-----	-----		
culture	s292		s292	s292	science	-----	-----		
darkness	-----	-----			score	-----	-----		
defense	-----	-----			society	s300		s300	s300
demand	-----	-----			space	-----	-----		
discovery	-----	-----			speed	-----	-----		
effort	-----	-----			sucess	-----	-----		
empire	-----	-----			taste	-----	-----		
feature	-----	-----			tax	-----	-----		
folk	-----	-----			territory	-----	-----		
force	s312		s312	s310	text	-----	-----		
frequency	-----	-----			thing	-----	-----		
friend	-----	-----			trial	-----	-----		
game	-----	-----			tribe	-----	-----		
goal	s4		s4	s4	type	-----	-----		
government	-----	-----			unit	-----	-----		s672
greatness	-----	-----			value	-----	-----		
hand	-----	-----			variety	-----	-----		
head	-----	-----			way	-----	-----		
history	-----	-----			wiiner	-----	-----		
importance	-----	-----			world	-----	-----		
impulse	s19		s19	-----					
knowledge	-----	-----							
measure	-----	-----							
nation	-----	-----							
objective	s14		s14	s14					

Table A-5.6 Experiment 2 - first and fifth paragraph

	dall-p0	dall-p4	d0X-3		dall-p0	dall-p4	d0X-3
Interpretation	#1	#1	#1	Interpretation	#1	#1	#1
Score	15	0	18	Score	15	0	18
token				token			
achievement	-----	-----		opponent	-----	-----	
aggression	-----	-----	s245	order	-----	-----	
area	-----	-----		path	-----	-----	
arts	-----	-----		people	-----	-----	
aspect	-----	-----		person	-----	-----	
binding	-----	-----		place	-----	-----	
book	-----	-----		population	-----	-----	
case	-----	-----		power	-----	-----	
cash	-----	-----		problem	-----	-----	
challenge	-----	-----		production	-----	-----	
change	-----	-----		purpose	-----	-----	
chapter	-----	-----		range	-----	-----	
civilization	s693		s693	rate	-----	-----	
component	-----	-----		rear	-----	-----	
conquest	-----	-----		reference	-----	-----	
course	-----	-----		rest	-----	-----	
culture	s292		s292	science	-----	-----	
darkness	-----	-----		score	-----	-----	
defense	-----	-----		society	s300		s300
demand	-----	-----		space	-----	-----	
discovery	-----	-----		speed	-----	-----	
effort	-----	-----		success	-----	-----	
empire	-----	-----		taste	-----	-----	
feature	-----	-----		tax	-----	-----	
folk	-----	-----		territory	-----	-----	
force	s312		s312	text	-----	-----	
frequency	-----	-----		thing	-----	-----	
friend	-----	-----		trial	-----	-----	
game	-----	-----		tribe	-----	-----	
goal	s4		s4	type	-----	-----	
government	-----	-----		unit	-----	-----	
greatness	-----	-----		value	-----	-----	
hand	-----	-----		variety	-----	-----	
head	-----	-----		way	-----	-----	
history	-----	-----		winner	-----	-----	
importance	-----	-----		world	-----	-----	
impulse	s19						
knowledge	-----	-----					
measure	-----	-----					
nation	-----	-----					
objective	s14		s14				

Table A-5.7 Experiment 2 - first and sixth paragraph

	dall-p0	dall-p5	d0X-4		dall-p0	dall-p5	d0X-4
Interpretation	#1	#1	#1	Interpretation	#1	#1	#1
Score	15	11	31	Score	15	11	31
token				token			
achievement	-----	s455	s455	opponent	-----	-----	
aggression	-----	-----		order	-----	-----	
area	-----	-----		path	-----	-----	
arts	-----	s267	s267	people	-----	-----	
aspect	-----	-----		person	-----	-----	
binding	-----	-----		place	-----	-----	
book	-----	-----		population	-----	-----	
case	-----	-----		power	-----	-----	
cash	-----	-----		problem	-----	-----	
challenge	-----	-----		production	-----	-----	
change	-----	-----		purpose	-----	-----	
chapter	-----	-----		range	-----	-----	
civilization	s693	s693	s693	rate	-----	-----	
component	-----	-----		rear	-----	-----	
conquest	-----	-----		reference	-----	-----	
course	-----	-----		rest	-----	-----	
culture	s292	s292	s292	science	-----	-----	
darkness	-----	-----		score	-----	-----	
defense	-----	-----		society	s300	s300	s300
demand	-----	-----		space	-----	-----	
discovery	-----	-----		speed	-----	-----	
effort	-----	s318	s318	success	-----	-----	
empire	-----	-----		taste	-----	-----	
feature	-----	-----		tax	-----	-----	
folk	-----	-----		territory	-----	-----	
force	s312	-----	s312	text	-----	-----	
frequency	-----	-----		thing	-----	-----	
friend	-----	-----		trial	-----	-----	
game	-----	-----		tribe	-----	-----	
goal	s4	-----	s4	type	-----	-----	
government	-----	-----		unit	-----	-----	
greatness	-----	-----	s326	value	-----	-----	s272
hand	-----	-----		variety	-----	-----	
head	-----	-----		way	-----	-----	
history	-----	s287	s287	winner	-----	-----	
importance	-----	-----	s22	world	-----	-----	
impulse	s19	-----	s19				
knowledge	-----	-----					
measure	-----	-----					
nation	-----	-----					
objective	s14	-----	s14				

Table A-5.8 Experiment 2 - first and seventh paragraph

	dall-p0	dall-p6	d0X-5		dall-p0	dall-p6	d0X-5
Interpretation	#1	#1	#1	Interpretation	#1	#1	#1
Score	15	0	17	Score	15	0	17
token				token			
achievement	-----			opponent	-----		
aggression	-----			order	-----		
area	-----			path	-----		
arts	-----			people	-----		
aspect	-----			person	-----		
binding	-----			place	-----		
book	-----			population	-----		
case	-----			power	-----		
cash	-----			problem	-----		
challenge	-----			production	-----		
change	-----			purpose	-----		
chapter	-----			range	-----		
civilization	s693		s693	rate	-----		
component	-----			rear	-----		
conquest	-----			reference	-----		
course	-----			rest	-----		
culture	s292		s292	science	-----		
darkness	-----			score	-----		
defense	-----			society	s300		s300
demand	-----			space	-----		
discovery	-----			speed	-----		
effort	-----			success	-----		
empire	-----			taste	-----		
feature	-----			tax	-----		
folk	-----			territory	-----		
force	s312		s312	text	-----		
frequency	-----			thing	-----		
friend	-----			trial	-----		
game	-----			tribe	-----		
goal	s4		s4	type	-----		
government	-----			unit	-----		
greatness	-----			value	-----		
hand	-----			variety	-----		
head	-----			way	-----		
history	-----			winner	-----		
importance	-----			world	-----		
impulse	s19		s19				
knowledge	-----						
measure	-----						
nation	-----						
objective	s14		s14				

Table A-5.8 Experiment 2 - first and eighth paragraph

	dall-p0	dall-p7	dall-p7	dall-p7	d0X-6	d0X-6		dall-p0	dall-p7	dall-p7	dall-p7	d0X-6	d0X-6
Interpretation	#1	#1	#2	#3	#1	#2	Interpretation	#1	#1	#2	#3	#1	#2
Score	15	4	4	4	16	16	Score	15	4	4	4	16	16
token							token						
achievement							opponent						
aggression							order						
area							path						
arts							people						
aspect							person						
binding							place						
book							population						
case							power						
cash							problem						
challenge							production						
change							purpose						
chapter							range						
civilization	s693				s693	s693	rate						
component							rear						
conquest							reference						
course		s458	s464	s464	s458	s464	rest						
culture	s292						science						
darkness							score		s475	s475	s468	s475	s475
defense							society	s300				s300	s300
demand							space						
discovery							speed						
effort							success		s444	s444		s444	s644
empire							taste						
feature							tax						
folk							territory						
force	s312				s312	s312	text						
frequency							thing						
friend							trial						
game				s684			tribe						
goal	s4				s6	s6	type						
government							unit						
greatness							value						
hand							variety						
head							way		s386	s195	s195	s386	s195
history							winner						
importance							world						
impulse	s19				s19	s19							
knowledge													
measure													
nation													
objective	s14				s14	s14							

Table A-5.9 Experiment 2 - first and ninth paragraph

	dall-p0	dall-p8	dall-p8	dall-p8	d0x-7		dall-p0	dall-p8	dall-p8	dall-p8	d0x-7
Interpretation	1	1	2	3	1	Interpretation	1	1	2	3	1
Score	15	16	15	15	37	Score	15	16	15	15	37
token						token					
achievement	-----					objective	s14				s14
aggression	-----					opponent	-----				
area	-----					order	-----	-----	s299	s299	-----
arts	-----					path	-----				
aspect	-----					people	-----				
binding	-----			s637	s627	person	-----				
book	-----	s650	s651	s654	s650	place	-----	s575	s575	s575	s575
case	-----					population	-----				
cash	-----					power	-----				
challenge	-----					problem	-----	s559	s559	s559	s559
change	-----	s617	s617	s617	s617	production	-----				
chapter	-----	s596	s594	s594	-----	purpose	-----				
civilization	s693				s693	range	-----				
component	-----					rate	-----				
conquest	-----					rear	-----	s496	s496	s496	s496
course	-----					reference	-----	s665	-----	-----	s665
culture	s292				s292	rest	-----				
darkness	-----					science	-----				
defense	-----					score	-----				
demand	-----					society	s300				s300
discovery	-----					space	-----				
effort	-----					speed	-----				
empire	-----					success	-----				
feature	-----					taste	-----				
folk	-----					tax	-----				
force	s312				s312	territory	-----				
frequency	-----					text	-----	s659	s659	-----	s657
friend	-----					thing	-----	s628	s628	s628	s624
game	-----					trial	-----				
goal	s4				s4	tribe	-----				
government	-----					type	-----				
greatness	-----					unit	-----				-----
hand	-----					value	-----				
head	-----	s517	s517	s517	s517	variety	-----				
history	-----					way	-----				
importance	-----					winner	-----				
impulse	s19				s19	world	-----				
knowledge	-----					world	-----				
measure	-----										
nation	-----										

Table A-5.10 Experiment 3 - first six paragraphs

Interpretation	d0-6_7_8-0	Sum of individual paragraphs 1-5	Interpretation	d0-6_7_8-0	Sum of individual paragraphs 1-5
	#1			#1	
Score	61	33	Score	61	33
token			token		
achievement	s455	s455	objective	s14	s14
aggression	s245		opponent		
area	s57	s57	order		
arts	s267	s267	path	s194	
aspect			people	s291	
binding			person		
book			place		
case			population	s186	
cash			power	s145	s145
challenge			problem		
change			production		
chapter			purpose	s109	
civilization	s693	s693	range	s223	
component			rate	s108	s107/s108/s105
conquest			rear		
course			reference		
culture	s292	s292	rest		
darkness	s40		science	s181	s181
defense			score		
demand			society	s300	s300
discovery	s198		space		
effort	s318	s318	speed	s114	s114/s112
empire			success		
feature			taste	-----	
folk			tax	-----	s178
force	s314	s312	territory	s44	s44
frequency	s182		text		
friend			thing		
game			trial		
goal	s4	s4	tribe		
government			type		
greatness	s326		unit	-----	
hand			value	s273	
head			variety	s89	
history	s287	s287	way	s391	s386/s195
importance	s22		winner		
impulse	-----	s19	world	s696	
knowledge	s362				
measure					
nation					

Table A-5.11 Experiment 4 - all paragraphs

Interpretation Score	dall-0 #1	Senses with no relations in this interpretation	Sum of individual paragraphs : 0-6 + 7 + 8	Interpretation Score	dall-0 #1	Senses with no relations in this interpretation	Sum of individual paragraphs : 0-5+ 6 + 7 + 8
	116		119		116		81
token				objective	-----		s14
achievement	s455		s455	opponent	s226		
aggression	s245			order	-----		s299
area	s57		s57	path	s195		
arts	s267		s267	people	s291		
aspect	s330			person	s532		
binding	-----		s637	place	s582		s575
book	s650		s650/s651/s654	population	s186		
case	s72			power	s145		s145
cash	-----	(s363)		problem	s559		s559
challenge	s454			production	-----	(s101)	
change	s617		s617	purpose	-----		
chapter	-----		s596/s594	range	s223		
civilization	s693		s693	rate	s108		s107/s108/s105
component	s431			rear	-----		s496
conquest	s207			reference	s665		s665
course	s464		s458/s464	rest	s588		
culture	s294		s292	science	s181		s181
darkness	s40			score	s475		s475/s468
defense	s238			society	s300		s300
demand	s156			space	s402		
discovery	s198			speed	s114		s114/s112
effort	s318		s318	sucess	s444		s444
empire	s162			taste	s211		
feature	s52			tax	-----		s178
folk	s491			territory	s44		s44
force	s314		s312	text	s657		s659
frequency	-----	(s182)		thing	s628		s628
friend	s480			trial	s545		
game	-----		s684	tribe	-----	(s62)	
goal	s6		s4	type	s565		
government	s679			unit	s669		
greatness	s326			value	s273		
hand	s350			variety	s89		
head	s517		s517	way	s195		s386/s195
history	s287		s287	winner	-----	(s340)	s386/s195
importance	s22			world	s696		
impulse	-----		s19				
knowledge	s362						
measure	-----	(s274)					
nation	s449						

